Community Structure in Endorsement Social Networks

Guillermo Garrido Yuste

Departamento de Lenguajes y Sistemas Informáticos

ETS. de Ingeniería Informática - UNED

Septiembre de 2010

Departamento de Lenguajes y Sistemas Informáticos
ETS. de Ingeniería Informática - UNED
Septiembre de 2010

Memoria de Tesis de Máster
Máster en Tecnologías del Lenguaje en la Web

Community Structure in Endorsement Social Networks

Guillermo Garrido Yuste

Curso 2009-2010

Director del Trabajo: Dr. Anselmo Peñas
Departamento de Lenguajes y Sistemas Informáticos
ETS. de Ingeniería Informática - UNED

Acknowledgements - Agradecimientos

Part of the work reported in this thesis was done in close collaboration with Aristides Gionis and Francesco Bonchi while I was visiting Yahoo! Research Barcelona. Many of the ideas in this report belong to them. Any error, mistake, or omission is on the contrary only attributable to me. The code for core selection and merging was produced by Aris Gionis. I also thank Carlos Castillo for helpful discussions and comments while visiting Yahoo! Research, and Ricardo Baeza-Yates, head of this institution.

Estoy muy agradecido al Dr. Anselmo Peñas, que me ha guiado desde mi comienzo en el Máster en Tecnologías del Lenguaje en la Web, hasta la entrega de éste, su trabajo final. Y quiero extender el agradecimiento al Departamento de Lenguajes y Sistemas Informáticos, dirigido por M. Felisa Verdejo, en el que he tenido la fortuna de trabajar los últimos dos años.

He contado para poder llevar a cabo este trabajo con la subvención del Ministerio de Ciencia e Innovación, a través del proyecto QEAVis-Catiex (TIN2007-67581-C02-01).

Contents

Abstract

In this work, we study the community structure of *endorsement networks*, i.e., social networks in which a directed edge $u \rightarrow v$ is asserting an action of support from user u to user v. Examples include scenarios where a user u is *favouring* a photo, *liking* a post, or *following* the microblog of user v. Very often, endorsement networks are sub-networks of more complex social systems; for instance, a photo-sharing site typically includes a "favouring" function, which induces an endorsement network. We start from the hypothesis that the footprint of a community in an endorsement network is a bipartite directed clique from a set of followers to a set of leaders, and apply frequent itemset mining techniques to discover such bicliques. Our analysis of real networks indicated that, with high statistical significance, this hypothesis holds, and that the leaders of a community are endorsing each other forming a *very dense nucleus*.

Our method produces many similar bicliques, which are different footprints of the same community. Thus, we propose a novel clustering technique in order to coalesce similar bicliques into meaningful communities. We explore different similarity measures based on set similarity and on edge density between followers and leaders, and by expressing edge density as an inner product operation we show how to make the clustering algorithm scalable. Our experiments demonstrate that our clustering algorithm is capable of discovering communities characterised by a set of leaders who link to each other and followers who link to the leaders.

True, false and other kinds of news radiated through the dormitory from these dense clusters.

White Noise, Don DeLillo

1 Introduction

The recent explosion of social on-line networking has created a variety of social-media services with many different purposes: connecting with friends, sharing multimedia content, entertaining, blogging, bookmarking, etc. (see section 2.3). For example, the popular social network site Facebook[1], which primary focus is on creating and maintaining a network of friends or acquaintances, reports having more than 500 million active users as of August 2010 [1]; the media sharing service Flickr[2] claimed to host 4 billion images in October 2009 [57]; and Twitter[3], the popular microblogging system, has recently revealed having over a hundred million users, who send an average total of 55 million messages a day [209].

There is a general request for methods and tools for analysing this wealth of social information:

> "From retailing to counter-terrorism, the ability to analyze social connections is proving increasingly useful." [208].

Understanding the viral spread of information in social media, modelling how information propagation relates to the underlying community structure, and identifying influential users, are some of the important challenges in social-network analysis. To give just a few examples, the use of social networking by the Obama campaign in the 2008 U.S. presidential campaign has been considered to play an important role in mobilising voters, and gathering record donations [222, 61] for the Democrat candidate; indeed, Twitter has been shown to be an indicator of political sentiment [211], and useful for advertising campaigns [143, 114], and even to forecast the success of products such as movies [23].

1.1 Endorsement Networks

As a step in the direction of understanding information propagation and identifying influential users, we have studied the community structure of what we call *endorsement networks*, in which a directed edge $u \to v$ is asserting a unit of support from user u to user v:

> We denote an endorsement network by $G = (V, E)$, where V is a set of nodes and E is a set of *directed* edges. A directed edge $(u, v) \in E$ indicates an action of endorsement from node u to node v.

[1] http://www.facebook.com
[2] http://www.flickr.com
[3] http://www.twitter.com

Two relevant examples of such endorsement networks come in the form of social media sharing platforms and microblogging services. These, and other on-line social network sites, are described in more detail in section 2.3.

In the popular media sharing system Flickr, a user u may *comment* or *favour* a photo of another user v. In case that u admires v's photos and wants to be updated on v's future posts, u may add v as a *contact*. In Flickr, contacts are unilateral, not necessarily symmetric, and they represent endorsement, not friendship. On the other hand, when a user u declares another user v as *friend* or *family*, the reason is that u wants to share her photos with v, and therefore this link represents social affinity rather than endorsement.

In microblogging services such as Twitter, users post short messages which are displayed on their profile page and delivered to the author's subscribers who are known as *followers*. Being a follower is an explicit form of endorsement. In some cases a user u might particularly like a post of user v and "retweet" it, thus propagating the content created by v. Twitter users can also send messages directly to each other. While direct messaging would imply a social acquaintanceship, the "following" relationship (endorsement in our terminology) has a more acute informational aspect. Java et al. [116] early study on the motivations for microblogging categorizes the intentions of users of such services in: information providers, information seekers and friendship-wise relationships. Moreover, Weng et al. [221] confirmed that the "following" relationship in Twitter is correlated with topical affinity. Thus, communities in such networks are not purely friendship-based, they are also based on common interests.

1.2 Why Endorsement Networks?

In this work, we use the term *endorsement network* to refer to directed social networks where an edge represents support from a user to another. We could extend the concept to consider networks that link users to *items*, and edges representing "liking" of the items by the users.

In section 2.2, we describe different kinds of social networks, according to the type of ties that link individuals. The usefulness of the term "endorsement" is that it stresses the difference between purely friendship-based social networks (that are undirected, as friendship is reciprocal), and networks where the direction of the edges is a fundamental property, with semantics of "support" from an user to another.

There is a need for better methods for understanding the spread of information and influence in this kind of networks. Weng et al. [221], for instance, argue for the need of better measures of influence than the simple in-degree.

The concept of endorsement networks is, then, founded in two reasons: the observation, tested in the existing literature, that in directed networks the directionality of the edges is a crucial property, not to be disregarded; and the intuition that, for a class of networks, the directions of the edges has a meaning of support from some users to others, and therefore measures of social influence in the network have to be designed with this kind of networks in mind.

The datasets we have worked with, belonging to this kind of networks and also to friendship-based social networks, are analyzed in section 3.2.

Analysing endorsement networks and understanding their community structure can lead to deeper insights in the the leaders-followers relationship, and ultimately, to mastering how information and user-generated content is propagating.

1.3 Community Structure of Endorsement Networks

In this work we study the community structure of endorsement networks, and the problem of detecting influential elements of the network. While community detection in social networks is a very well studied problem in statistical physics, sociology and computer science (see section 2.5), here we are interested in particular in the community structure of social endorsement networks, where the links between nodes have different semantics than in usual social networks. In addition, the community-detection task in endorsement networks has other peculiarities: links are directed and communities are naturally overlapping. More precisely, we expect "semi-overlapping" communities, where each community has a group of leaders strongly mono-thematic, while most of the users are naturally multi-thematic and thus following different groups of leaders.

As we will see in section 2.5, most variants of the community detection problem are NP-hard, and sometimes even the task of approximate solving cannot be completed in polynomial time. A large number of heuristics have been proposed for this family of problems, but the problem is not yet satisfactorily solved [84]. On the other hand, even methods with polynomial but super-linear complexity do not scale enough to be usable for the very large networks that recent research, specially on Social Network Sites (see 2.3), aims to tackle. As the data we are most interested in comes from on-line social network services, we want methods capable of dealing with very large datasets.

The task of community detection in directed networks has additional difficulties. The most common approach in the past literature has been to disregard directions in the network, and use methods designed for undirected graphs. It has been shown that this methodology can give strange results and counter-intuitive communities [141]. A good community detection method should take into account the directionality of the edges.

In our view, in real-world social networks, a clear-cut partition of a network in disjoint groups of nodes is not a realistic depiction of its community structure. In endorsement networks, we assume that a user can be strongly interested in more than one topic, and can be part of more than one "community". A good method for community detection in endorsement networks should allow for overlapping communities.

For some networks, a hierarchical community structure has been observed: groups of nodes can be merged to form larger groups. A method that delivers some form of hierarchy would be useful to uncover this kind of structures.

Our contribution aims at discovering communities in very large networks, rather than partitioning the entire graph in subgroups. And within communities, a group of influential elements, or *leaders*, and the elements the influence, or *followers,* are singled-out.

1.4 Research Goals

On a high level, this investigation has two main general goals:

1. To study the community structure of very large endorsement networks within online social network services. Starting from a thorough empirical study of examples of these networks, we want to devise a successful method to discover these communities.

2. To obtain insights into the spread of influence in these networks.

These goals are expressed through the following specific goals:

1. Propose a successful *method for community detection in endorsement networks.* Such a method should have the following properties:

   - It should be capable of dealing with very large networks, as the real-world applications of the methods we propose should scale up to the very large size of social network sites. This goal suggests that methods that use only local information of the network will be better suited.
   - We are interested in endorsement networks: therefore we want to take into account, in a natural way, the directions of the edges.
   - We assume that realistic communities in these networks have a significant overlap; so the method should allow for the natural overlap of communities in real-world networks.

   Our hypothesis is that it is possible to design such a method if we look for certain topological substructures of the network that are strong indicators of community structure. These structures would be the *footprint* of communities; to decide on a suitable footprint, we start from a thorough empirical analysis of real-world networks.

2. A second specific goal is to obtain measures of influence in endorsement networks. We argue that both problems are tightly entwined: our insight is that influential individuals, *leaders*, are so within a community; and that *leaders* have to be defined in relation to those other individuals that more closely *follow* their activities.

3. A secondary intention of the work performed is to quantitatively assess the differences between endorsement networks and friendship-based social networks.

1.5 Outline of the Investigation

The research that is the subject of this thesis report has been empirical, and guided from experimental investigation. We started from a thorough examination of real-world networks, and drew from our empirical observations the insights that we have tried to follow in the implementation of the methods proposed.

We draw the first cue to guide our investigation from the work by Kumar et al. [132] (see section 2.5.2). They suggest that communities of pages in the Web are characterised by dense directed bipartite subgraphs. Furthermore, they hypothesise and test on Web data that any large enough bipartite directed subgraph of the Web almost surely has a *core*. Such a core is a complete bipartite subgraph. Recall that, for a bipartite subgraph formed by node sets A and B to be complete, every possible link from nodes in A to nodes in B must be present. Cores can also be called bipartite cliques, or *bicliques*.

Therefore we started from the hypothesis that such bicliques from followers to leaders can be found also in a social endorsement network. In order to find such bicliques in a large endorsement network, we apply *frequent itemset mining* techniques to an adjacency-list representation of the network: we consider each edge in the network to be a transaction, which represents the outgoing neighbours of the node.

As it is usually the case when mining any form of frequent patterns, our method produces many similar and redundant cores, which presumably are different footprints of the same community. Thus we propose a novel clustering technique in order to coalesce similar cores into meaningful communities. For the clustering algorithm we need to define a measure of similarity between cores. We explore different alternatives that rely on set similarity and on edge density between followers and leaders of the cores. As a technical contribution, we show how to express the edge density measure as an inner product of appropriately defined vectors, and therefore obtain significant computational gains.

Through our analysis of real-world endorsement networks we have discovered:

- **Large cores with very dense nuclei:** endorsement networks contain large bicliques from a set of followers to a set of leaders. The set of leaders (nucleus) of a core almost always exhibits an extremely high internal density.

- **Communities:** by coalescing cores we find a reasonably small number of communities having a very large followers base, while still maintaining a very high density in their leadership nucleus.

1.6 Organization of this Thesis

The structure of this work is organized as follows. In Chapter 2, the background and related work is presented. We first define a few fundamental graph theoretical concepts (section 3.1); then, examples of real-world networks are presented (section 2.2). Our experiments are performed with data gathered from on-line social network sites, so we discuss them in detail in section 2.3. The methodology of social network analysis is analyzed in section 2.4. An important section of this related work chapter is section 2.5, where we describe the phenomenon of communities in networks and review the state of the art of community detection methods. In section 2.6, a few concluding remarks help situating our work in the background.

Chapter 3 reports on the approach, methodology and results of our empirical investigation. The working definitions are collected in section 3.1. The datasets used are described and analyzed thoroughly in section 3.2. Our experiments are reported in sections 3.3 and 3.4. The results obtained are discussed in section 3.6, in relation to the research goals we describe above.

Finally, Chapter 4 summarizes what we have learnt from this work, and the research questions that remain open for the future.

2 Background and Related Work

2.1 Graphs

In this section, we define some of the basic concepts we will use throughout this work.

A ***graph***, or ***network***, is a mathematical representation of the relationships among a set of objects. A graph G consists of a set of ***nodes***, V, and a set E of links between pairs of nodes, called ***edges***: $G = (V, E)$[1].

Two important quantities of a graph are the number of nodes and the number of edges, in the following, we will use the letter n for the number of nodes ($n = |V|$) , and m for the number of edges ($m = |E|$).

The edges of a graph can be directed or undirected. A directed edge represents a relation of the kind u points to v, while an undirected edge represents a reciprocal relationship: u points to v, and v also points to u.

A ***directed graph*** consists of a set of nodes and a set of directed edges. In a directed graph, the direction of the links is an important property. An ***undirected graph*** is a graph where all edges are reciprocal: it consists of a set of nodes and a set of undirected edges. More formally, in an undirected graph, the set of edges is a set of unordered pairs, whereas for a directed graph it is a set of ordered pairs. An undirected graph can equivalently be represented as a directed graph where for every edge its reverse is also an edge of the graph[2].

Graphs can have *labels* on the edges; in particular, for numerical weights, we have *weighted graphs. A* ***weighted graph*** gives a numerical weight to every edge. More formally: a weighted graph $G = (V, E, W)$ consists of a set of nodes, V, a set of edges, E, and a set of weights, such that each edge has an associated weight. Weights can be any real number, but it is common to have graphs where weights are restricted to be rational, integer, or positive integer numbers.

Two nodes that have an edge between them are called ***adjacent***. If an edge e has node v as one of its end-points, we say that e is ***incident*** to v.

A ***loop*** is an edge that has the same node in both its endpoints, that is: $l = (v, v)$. An undirected graph with no loops and no multiple edges is called a ***simple graph***.

A useful way of representing mathematically a graph is the ***adjacency matrix***. If G is a graph with n nodes, $v_1, v_2, \ldots, v_n$, its adjacency matrix is $A = (a_{ij})_{i,j=1\ldots n}$ where a_{ij} is the number of edges from node v_i to node v_j. We are usually interested in simple graphs,

[1] In this work, we prefer the terms *node* and *edge*. In mathematics, nodes are commonly called *vertices*; in social sciences, the term *actor* is frequently used. Edges are often called *links* in the context of computer science, or *ties* in sociology.

[2] A directed edge is often called *arc*, and a directed graph is frequently called a *digraph*.

and for them, the adjacency matrix is $A = (a_{uv})_{u,v \in V}$, where $a_{uv} = 1$ if $(u, v) \in E$ and $a_{uv} = 0$ otherwise.

A ***subgraph*** S of a graph $G = (V, E)$ is a subset of the nodes, and the edges between them:

$$S = (U, F), \text{ where } U \subset V \text{ and } F = \{(u, v) \text{ where } u, v \in U\}$$

Graphs are an abstract representation that can be used to model any set of objects and the relationships between them: a network (in the literature, graph and network are terms used synonymously). This abstraction has been used in multiple domains; in section 2.2, we show very different examples of networks.

We introduce now a few fundamental graph-theoretical concepts.

Paths Given a graph, $G = (V, E)$, a ***path***, p, from node u to node v, is a sequence of nodes, $A = v_1, v_2, ..., v_k = B$, where each pair of consecutive nodes have a link between them: $(v_i, v_{i+1}) \in E$, for each $i = 1, ..., k - 1$. A path describes a way of reaching a node from another node, by following the edges of the graph.

If the graph is directed, a ***directed path*** is a path that respects the direction of the edges. Unless specified otherwise, if we refer to a path in a directed network, we are referring to a directed path.

A ***simple path*** is a path were nodes appear at most once. A ***cycle*** is a path, of three nodes or more, where first and last node are the same, and all other nodes are different. A graph is ***acyclic*** if it contains no cycles.

Distance The ***length*** of a path is the number of edges it contains. The length of the path $u = v_0, v_2, ..., v_k = v$ from u to v is k. The ***distance*** (u, v) between two nodes of a graph, u and v, is the length of the *shortest path* between them, if any path exists. Such shortest path is often called ***geodesic path***.

We can define the distance from any node to itself as 0: $d(u, v) = 0$; and the distance between two nodes where no linking path exists as infinity: ∞.

The ***diameter*** of a graph is the largest distance between a pair of nodes. In other words, it is the largest shortest path between any two vertices in the graph[3].

Degree The ***degree*** $d(v)$ of a vertex v in a graph G is the number of nodes adjacent to v. Equivalently, it is the number of edges incident to v (loops from v to v are counted twice).

In a directed graph, we can distinguish two definitions of degree. The ***in-degree*** of node v is the number of nodes that point to v: $d_{in}(v) = |\{(u, v) : u \in V\}|$. The ***out-degree*** of node v is the number of edges starting at v: $d_{out}(v) = |\{(v, u) : u \in V\}|$.

[3] Some authors define diameter as the average shortest path in the graph.

Triangles, cliques, and clustering coefficient An interesting property of real world graphs is the existence of triangles. A ***triangle*** is a set of three nodes, u, v, and w, such that there are edges connecting each other[4].

A ***clique*** in a graph is a subgraph where any pair of nodes is adjacent. A triangle is a clique with three nodes. Since a graph where every pair of nodes is connected by an edge is called a ***complete graph***, a clique is a complete subgraph of a larger graph G.

In real-world networks, triangles appear frequently: there is a high ***transitivity*** in the network, in the sense that if nodes u and v are neighbours of a third node, w, then it is likely that u and v are neighbours themselves. The ***clustering coefficient***, C, is a measure of this characteristic of networks: it measures the degree to which the neighbours of a particular node are connected to each other[5] [41]. Two definitions of the clustering coefficient for undirected networks have been proposed in the literature:

The first defines the clustering coefficient of the complete network. A connected subgraph of three nodes is called a *connected triple*. If the 3-node subgraph is complete, it is a *triangle*; we want to measure how many of the connected triples are in fact triangles. So we define the clustering coefficient as the proportion of connected triples in the network that are in fact triangles:

$$C = \frac{3(number\, of\, triangles)}{number\, of\, connected\, triples}$$

Clearly, the factor 3 is necessary because each connected triple is counted once for each of its three nodes. The number of connected triples a node participates in can be computed as the number of combinations of pairs of its neighbours, and therefore, for undirected graphs:

$$number\, of\, connected\, triples = \sum_{v \in V} \frac{d(v)(d(v)-1)}{2}$$

where $d(v)$ is the degree of node v.

This definition of the clustering coefficient corresponds to the concept of fraction of transitive triples used in sociology [41].

An alternative definition was introduced by Watts and Strogatz [219]. Rather than for the entire graph, this clustering coefficient is a measure of the local transitivity around a single node. The clustering coefficient of node v is the probability that two randomly selected adjacent nodes of v are also adjacent. Let node v have degree $d(v)$ and let $e(v)$ be the total number of edges among all neighbours of v. The clustering coefficient of v is then defined as the ratio between $e(v)$ and the maximum possible number of edges among the neighbours of v:

[4]For an undirected graph, the definition is clear: nodes u, v, w form a triangle if there are three edges linking the three nodes: (u, v), (v, w) and (u, w).

In a directed graph, the definition of a triangle also requires the existence of three edges joining the nodes, but, depending on the directions of the edges, different kinds of triangles can be considered.

[5]Note that, despite its somewhat confusing name, the clustering coefficient is not related to *clustering* (unsupervised machine learning), and it is not a measure of the quality of a *clustering*.

$$C(v) = \frac{e(v)}{d(v)(d(v)-1)/2} = \frac{2e(v)}{d(v)(d(v)-1)}$$

Watts and Strogatz defined the clustering coefficient to be 0 for nodes with degree less than 2, and proposed as clustering coefficient for the whole graph the average of $C(v)$ over all the nodes of the network. Observe that the two definitions yield different values, although they measure the same phenomenon and have similar properties [41].

Connectivity in graphs A graph is ***connected*** if there exists a path between any pair of nodes. A graph that is not connected can be decomposed in connected subgraphs.

A ***connected component*** of a graph G is a maximal connected subgraph of G. In other words, it is a subset of the nodes such that for every pair of them there is a path from one to the other, and it is not contained in a bigger connected graph.

Any graph can be therefore partitioned in the set of its connected components (if the graph is itself connected, it will have just one component, the entire graph).

For directed graphs, two notions of connectivity have been used. A strong definition of connectivity requires a connected component to contain a *directed* path between any pair of nodes. A ***strongly connected component*** of a directed graph G is a maximal subset of nodes such that there is a directed path between any pair of nodes. So for nodes u and v, there must be a directed path from u to v and a directed path from v to u. We can also consider a weaker definition of connectivity. A directed graph is ***weakly connected*** if the underlying undirected graph[6] is connected.

Density If we consider a graph with no loops, and where multiple edges (edges having the same end-points) are not allowed, there is a maximum number of possible edges for a given number of nodes. The density is the proportion of those edges that the graph actually has.

For undirected graphs, density is defined as:

$$\delta = \frac{|E|}{|V|(|V|-1)/2} = \frac{2|E|}{|V|(|V|-1)}$$

A graph is said to be ***sparse*** if the number of edges is in the order of the number of nodes: $|E| = O(|V|)$; otherwise, it is called **dense**,

2.2 Real-world Networks

In this section, we show examples of networks from different domains, and relevant research devoted to the study of each one.

[6] The underlying undirected graph is the same graph, but ignoring directions: an undirected graph that has an edge whenever there was an edge, in any direction, in the directed graph.

2.2.1 Social networks

The nodes of a social network are people, or groups or people, and the edges represent interactions or contacts among them. Societies, and groups within them, form social networks.

Traditional social network research methodology is difficult and time consuming, as it is based on surveying and interviewing individuals from the network. In section 2.4.3, we discuss in more detail this kind of studies, and analyze further the advantages and drawbacks of this approach. Recent research has tried to use other methods to gather social network data. The success of Social Network Sites has provided researchers with a source of "hard data" on the contacts and interactions among people. In section 2.3, we review studies focused on this kind of data and discuss the difficulties that arise when using this methodology.

Relationships within a social network (commonly called *ties* in the social sciences literature) are of different nature: friendship, interaction, etc. Borgatti et al. [40] classify the kind of interactions that interest social network analysis in four types: similarities, social relations, interactions and flows[7]:

1. Similarities: edges in the network represent that the linked nodes are similar respect to some defined criterion. These kind of ties can be further classified in:
   a) location: for example being at the same place.
   b) membership: belonging to a certain group, or attending to the same event.
   c) attribute: having the same gender.
2. Social relations: this kind of ties represents that nodes have some kind of social connection. Some of them would be:
   a) kinship (family ties).
   b) social ties: friendship, *being employee of*, *being student of*, etc.
   c) affective: e.g., *likes*, *hates*.
   d) cognitive: e.g., *knows*.
3. Interactions: this group of ties represents social interactions between the nodes. Examples include: having talked; or having given advice to.
4. Flows: this last group would include ties like flows of information or resources. One of the basic intuitions of social network analysis is that the flow of information in a network depends basically on the social structure.

In the following, we will mention examples of relevant research on these different types of networks. We leave out those that we consider "information networks": those where nodes represent "information", rather than people. Although they often have a clear "social" component, we prefer to treat them separately in section 2.2.2.

[7] It is interesting to realize that, while physical scientists typically try to find out what networks with very different kinds of connections have in common, social scientists look for the differences between them, analytically and theoretically (see section 2.4.3).

Similarity social networks Similarity ties link social actors (people) that have some common feature. A **location** tie would represent that two people were in the same place at the same moment. An example of this kind of network has been used to explain the spread of a virus in a population. A contagious disease will have opportunities to spread through a *contact network* [74, p. 645], a network where a node is a person and an edge represents that two people came into contact in a way that makes the contagion possible. Examples of research on contact networks include studying how travelling behaviours affect the spread of a disease [77, 157, 60], or how interactions between animals[8] form so-called epidemic trees, that explain disease outbreaks [111]. In all these studies, the spatial location of the individuals of the network helps in explaining the spread of the disease.

A **membership** tie links two people who, for instance, belong to the same group. An example of affiliation network studied in the literature is the membership of people to the board of directors of major corporations [160].

An **attribute** tie links two people who share a common attribute, like gender, race, or attitude.

Social relations **Kinship** ties link members of a family. They are not the focus of interest of current social science, but graphical depictions of networks were first used to represent kinship relations. In the eighth century, trees were drawn to represent family structures [88].

Friendship patterns were already studied in the 1920s and 1930s by one of the pioneers of social network analysis, Jacob Moreno [162] and in the 1960s in the study of schoolchildren friendship networks by Rapoport et al. [188].

Relationships in the **workplace** were studied by the Harvard group of Mayo, who focused in Western Electric employees' productivity (see [88]). Sexual relationship networks have also received attention [149], who showed that this kind of networks are effectively "small worlds".

Examples of **affective** ties are relationships "A likes B" or "A hates B". In some Social Network Sites, a functionality that allows users to explicitly express this kind of relationships has been implemented. For instance, the *geek*-related news website Slashdot[9] implemented in 2002 a system that allowed users to mark other users are either "friends" or "foes". Leskovec et al. [146] studied models for predicting the sign of the links (positive or negative), and linked their findings to theories from social psychology.

Collaboration networks are a much researched source of data. Examples of collaboration graphs are the co-authorship network among scientists of a discipline[10], or the

[8] Nodes in these networks are animals and not *people*, but the analogy is clear.

[9] http://slashdot.org

[10] In the paper co-authoring network, nodes are scientists, and there is a link between two of them if they have collaborated in writing a paper. A peculiar subgraph of the paper co-authoring network is the connected component that contains the most collaborative mathematician of all time, and one of the most important figures of network research, Paul Erdős. Erdős was tireless, and collaborated with at least 458 other authors. This lead to the concept of *Erdős number*, defined to be 0 for Erdős himself, 1 for those who co-authored a paper with him, 2 for the collaborators of those, and so on [103]. It turns

network of collaboration of film actors[11].

Other relevant examples of collaboration networks are the Wikipedia collaboration graph, that connects editors of the same article and the network of collaboration of executives in a board of directors [67].

Interaction This kind of ties link individuals that interact in a social space. The "who talks to whom" networks are good examples. Social Network Sites provide researchers with large scale sources of data. The Microsoft IM graph (see section 2.4.2) is a very large snapshot of the conversations between users of that chat service over a month. Similar studies have been carried out on networks of emails [75, 106, 170, 4, 130], and phone calls [10, 11, 176].

2.2.2 Information networks

In an information network, the nodes represent information, rather than people (although examples like the citation network, below, have a clear social component).

Maybe the most prominent example of an information network is the World Wide Web (or simply, the Web). The Web can be modelled as a directed graph where the nodes are the HTML pages, and the edges are the hyperlinks that connect pages to each other, allowing for the now familiar experience of browsing. The Web graph was studied by Kleinberg, Kumar, and others [131, 129]. Understanding its structure is useful for various reasons [47]: to design better crawling strategies [58]; understand the sociological dynamics of Web content creation; or to predict the evolution of the Web. The hyperlink structure of the web has been exploited to design search algorithms [35, 178, 129], or to implement topic classifiers [56]. As we will discuss in more detail below it has also been mined for communities [132] (see section 2.5.2).

Research on the Web hyperlink structure was inspired by earlier works, specially on academic citation networks (see [125]). Nevertheless, both networks have different features, as the early work on the Web graph already pointed out [125, 178]. While academic publications are thoroughly reviewed, Web pages can be created freely and with negligible publication cost. Therefore, any algorithm on the Web may be a target of manipulation. On the other hand, quality, length and topics are much more diverse on the Web.

Garfield studied the network of citations among academic publications, and proposed a measure of the quality of a publication. His so-called science citation index results from dividing the number of citations received by the number of papers published in a certain period [89, 90].

out that science is indeed a "small-world" (see section 2.4.2): most mathematicians have an *Erdős number* smaller than 6, and scientists in other disciplines have comparable numbers: Albert Einstein's is 2, Noam Chomsky's is 4, Francis Crick's and James Watson's are 5 and 6 respectively [74].

[11] The actor collaboration network links actors that have appeared in the same film. Using the Internet Movie Database, three students adapted the idea of the Erdős number to compute the *Kevin Bacon number*: the length in the network of actor's collaborations, from Kevin Bacon to any other actor. It turns out this world is also very small, with an average Bacon number of 2.9 [74].

Another interesting example of network where nodes represent "knowledge" are the semantic word networks extracted from Wordnet, studied by Sigman and Cecchi, among others [200].

2.2.3 Natural networks

Networks in the natural world have recently been the centre of a great attention from the research community. Natural ecosystems have been represented by *food webs* [161, 92, 181, 223], graphs where the nodes represent species and the edges predator-prey relationships. It is also possible to derive *predator overlap graphs* where nodes are predators, linked if they share a common prey, or the analogue *prey overlap graphs* [181].

The network of neural connections has also been focus of attention. The nodes are neurones, and edges represent synaptic connections between two neurones [205]. This research has used the simple network of connections of the worm *C. Elegans*, that has 302 nodes and around seven thousand edges[2].

Another example from biology is the network mapping of the cell's metabolism [117, 26]. In this case, the nodes are the specific molecules or groups of molecules that interact in metabolic reactions, and the edges represent these chemical interactions.

In chemistry, networks have been used to study chemical reaction networks [16].

2.2.4 Technological networks

A classic example of a technological network is the electrical power grid [219, 18], where the nodes are electric substations, generators and transformers, and the edges represent transmission lines among them.

The Internet, the wired infrastructure of routers over which the World Wide Web is just one, if the most famous, application, has also been studied from the network analysis perspective [79]. The nodes are the computers ("routers" of the network), and the edges are the connections that link them[12].

2.3 Social Network Sites

Research has found in Web based services a source of data that is easier to gather, and arguably less affected by certain data-collection biases, than traditional, interview-based social network data. Our own work deals with this kind of data, so we find important to discuss this services in detail.

In this section, we analyze where Web-based social network data comes from, and the subtle differences between networks people take part in society, and networks within these services. First, we look at the general picture of services that can be described with the "umbrella terms" Web 2.0 and Social Media (section 2.3.1). Then, we discuss in detail the concept, history and research of an important "subclass": Social Network Sites

[12]It is interesting to observe that these physical networks are, on a higher level, "economic networks" [74]. The Internet, at a higher level, can be thought of as a "who-transacts-with-whom graph [...] that represents the data transfer agreements these Internet service-providers make with each other." [74].

(section 2.3.2). Important for our research are the features of some of the sites under this general description, so we dedicate section 2.3.3 to social media sharing, and section 2.3.4 to a form of "microblogging" of which Twitter is the most successful example.

2.3.1 Web 2.0 and social media

Easley and Kleinberg [74] point out three factors defining the enriched Web environment that has come to be described by the vague term of Web 2.0[13]:

1. New applications enabled users to generate and share content.
2. Users moved more and more of their personal data to web-based services.
3. If in the traditional design of the Web documents were linked, the new applications allowed to link *people*.

Easley and Kleinberg identify these three factors in a wide range of services: Wikipedia embodies factor (1); email services are based on factor (2); friendship-based services like MySpace or Facebook, at least primarily, are focused on creating and maintaining a social network: factor (3). Media sharing services like Flickr or YouTube join together the three factors: share photographs or videos (2), tag and comment on them (1), and have a social network of other users who they followed (3). Twitter, to give just another example, would be based on factors (1) and (2).

Social media researcher danah boyd[14] uses the umbrella term "social media" to describe the set of "tools, services, and applications that allow people to interact with others using network technologies" [63]. This is also a very broad term that would include any software-based medium that allows people interaction: from on-line communities to media-sharing, peer-to-peer networks and network gaming: "instant messaging, blogging, microblogging, forums, email, virtual worlds, texting, and social network sites are all genres of social media" [63].

Although many of these tools and services offer valuable data that can be mined to study the patterns of connectivity and the behaviour of users, in our work we are interested in Social Network Sites in particular.

2.3.2 Social Network Sites

The emergence in the past few years of Social Network Sites has provided researchers with reliable large scale data on the connections and interactions of people. Much of the current research in social networks takes advantage of the readily availability of data coming from these web services. Dually, the existence of these services, the way people interact on them, might be influencing the views on social networks of the users, and of researchers. The users might be more aware, or aware in different ways, of what is their network of acquaintances.

[13] A term suggested by Tim O'Reilly and others [177]

[14] We respect here the capitalisation (or rather, lack of it) this researcher has chosen for her name. See www.danah.org/name.html

Boyd and Ellison proposed a working definition of Social Network Site (SNS) as:

> "web-based services[15] that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system." [64]

Three elements are therefore important in the data exposed in SNSs in the context of social network analysis:

1. The user profiles. It is interesting to observe that SNSs have qualities of social network and also of information network. People indeed interact with each other, and manage a network of friends and acquaintances, like they would do in their "real-world" social network. They also maintain a profile, and interactions in the network also occur in the information level. There is a subtle difference with social networks outside these systems, where *people* interact with each other. For instance, in many popular SNSs, a profile might represent a company, an institution, a group of like-minded people and so forth[16].

2. The users with whom a user shares a connection. Multiple kinds of connection are possible. Boyd and Ellison's definition seems to be focused on "friends lists", like those maintained in Facebook or MySpace; but multiple kinds of connections are possible, within the same site. On Twitter, for instance, it is possible to consider the strong ties formed by people sending messages to each other, and the weak ties formed by people following each other.

3. The possibility of traversing the network.

Boyd and Ellison's definition has been criticised by Beer [32], who argued that this classification, in trying to be wide and inclusive, is in danger of being too broad. There would be sites where people are preoccupied by maintaining and extending their network of connections while, in others, users are involved in other activities, result of what a network of connections is created. It would be the case of sites like YouTube, or Flickr, where the user engages not only on maintaining a public profile, but also, and this activity might be more important, on sharing a piece of media, like a video clip or a photograph. What other users are drawn upon is the media, and not the profile behind it. In section 2.3.3, we look at these sites in more detail.

SNSs have produced a huge amount of interest from academy and industry. Researchers hope to gain insights in the ways users engage with those systems; and not only that, they would like to learn about the structure of human society from studying the structure

[15] Observe that sites like Facebook are now also accessible through dedicated applications, or *apps*, on mobile phones and other devices, rather than "in the Web". This recent trend has important consequences, but we will not discuss them here (see [19]).

[16] A thorough discussion of this topic is outside the scope of this research. Let us mention that it has been argued that profiles are "commodities both produced and consumed by those engaged with SNS" [32], and there is active research on issues related to how identity is shaped by SNS participants (see, for instance [65, 99]).

of SNS user-to-user relationships. But, are they the same? Boyd and Ellison argue that in fact they are not, and suggest calling "Friends", capitalized, those connections maintained within a SNS, making a distinction with real world "friends"[17]. It is clear that the friendship neighbourhood of a person in a SNS would often overlap their real-world friend neighbourhood, but they will not be identical.

Citing Easley and Kleinberg:

> "[It] is an interesting and fairly unresolved issue to understand how closely the structure of friendships defined in on-line communities corresponds to the structure of friendships as we understand them in off-line settings"[74]

Also, the term "friend", when dealing with the kind of connections that are made within a SNS, is problematic. It is too broad and simplifying, as there are indeed different kinds of friendships[18].

History and research of Social Network Sites

Research on Social Network Sites has followed two separate methodologies: on one side, surveys and interviews; on the other, large-scale quantitative studies of the huge datasets these sites provide. While social scientists performing studies based on interviews and surveys defend that their work gets a deeper understanding of the motivations and behaviour of people, researchers using large datasets argue that their studies are the only capable of uncovering the large-scale patterns exhibited by the network as a whole. In this section, we mention a few highlights in the short but intense history of SNSs, and point out to examples of the two lines of research.

A good historical account of the birth and evolution of SNSs is given by boyd and Ellison [64]. The first website that can be ascribed to this category is the now defunct SixDegrees.com. Back in 1998, it allowed users to maintain a profile, have a list of friends and browse these lists of friends. Although this features existed in various systems at the time, SixDegrees was the first site to combine them [64]. In 2000, the service closed down. Since then, other sites have included social network functionality in different ways.

[17] Boyd and Ellison suggest that the SNSs are mediated social spaces, but that their egocentric structure, where the individual is at the centre of their own community "mirror[s] unmediated social structures"[64]. In boyd's work, the concept of mediated "networked publics" (interactions within SNSs) is compared with "unmediated publics" (public space social interactions). Boyd defends that "Networked publics support many of the same practices as unmediated publics, but their structural differences often inflect practices in unique ways"[63].

An immediate critique, formulated by Beer[32], is whether such a thing as *unmediated* social structures exist: "it is hard to think of a life offline, particularly for what appear to be the engaged and switched on youth"[32]. On-line and offline worlds are tightly entwined (as boyd and Ellison acknowledge).

[18] Paul Adams and his team at Google Inc. interviewed people and asked them to draw their network of friends, and arrange them in groups. People tend to group their friends in 4 to 6 different groups. And after asking the interviewees to name the different groups of friends, people came up with many different names, only 85% of which contained the word "friends". See Paul Adams' presentation [5]. An empirical study of different kinds of friendship can be found in [203].

Another early example of SNS was LiveJournal, a blogging site that included a feature allowing users to list other users as friends, in order to follow their posts and also tune privacy settings. Danah boyd studied Friendster through interviews [62]. Liben-Nowell et al. [148] studied the connectivity patterns of LiveJournal profiles that contained a postal code, and described a model of the geographical distribution of friendship.

In 2002, one of the most successful early SNSs was created: Friendster. It acquired a great popularity, but a conjunction of technical and social difficulties and heavy competition damaged it [64]. A number of other sites tried to replicate its success, with varying outcomes. Sites targeting specific demographics appeared: for instance, LinkedIn or Xing focused on business relations. Google's Orkut failed to be popular in the US but succeeded in Brazil, and later in India, where it is still very strong.

The popularity of SNSs prompted media sharing services to include social features, and become SNSs themselves: examples include Flickr, Last.FM and YouTube. This is very relevant to our research, so we come back to it in section 2.3.3.

MySpace was possibly the first SNS that received, in July 2005, global mainstream media attention, when it was acquired by News Corporation. Before that, it had grown immensely, by attracting users from three different demographics: indie bands wanting to reach their audience, teenagers, and post-college urban people [64]. From then on, the competitiveness of the SNS marketplace has been ever growing.

Facebook launched in 2004 as a Harvard only social network site. Gradually, it opened up to other American universities, high schools, a few companies, and finally to the general public. Today, Facebook counts with a massive user base[19].

An early study on the role of Facebook in American campuses was performed by Ellison et al. [76]. They surveyed a sample of the Facebook users at Michigan State University, and described the motivations and uses of the service. They found that the service helped its users to maintain connections and stay in touch with people who left the college, and therefore they could not reach offline anymore. Rather than to meet new people, they used Facebook to "intensify and solidify relationships that started offline" [76]. We see how on-line and offline social spaces start to be tightly entwined. Recent research, co-authored by researchers employed by Facebook [49], supported the idea that on-line and offline social activities reinforce each other: there is a *feedback loop*. Another early survey-based study of Facebook is [207].

Golder et al. [100] quantitatively analyzed a large-scale Facebook dataset[20]. Golder et al. studied the temporal patterns of Facebook messaging and showed that the average number of friends was[21] around 180. This value agrees with Ellison et al. study [76], and with the theoretical cognitive limit to the number of people with whom a person can maintain stable social relationships, proposed by the anthropologist Dunbar [73]. They also showed how, while most messages where sent to friends, most friend pairs do not exchange any message. This observation agrees with the existence of strong ties (friends

[19] According to statistics provided by Facebook, more than 500 million active users [1].

[20] Their dataset recorded the exchanging of 362 million messages over the span of almost two years by 4.2 million users[100].

[21] With a median of 144. The different in the two values is due to a small number of outliers: members with a very large number of friends (11 users had more than 10,000 friends) [100].

who message each other), and weak ties (those friends who do not) in the network. Researchers working at Facebook, have also shown how the network of friends does not match the network of communication [153]. They have suggested three (overlapping) categories of links, after studying a month's worth of user interactions:

- A *reciprocal communication link* if both users sent messages to each other.
- A *one-way communication link* represents a message was sent from an user to another, irrespective of whether it was reciprocated.
- A *maintained relationship link* if the user followed information from the other (by, for instance, browsing the friend's profile), whether or not they exchanged messages.

The results of this research supports the idea that it is possible to map, over the same set of users, different networks with different strengths: weak (there is a friendship contact), or strong (actual communication, through messages). Marlow et al. add a new kind of ties, in-between strong and weak: "maintaining relationship with". These are only possible *because* the technology of SNSs enables them: it is the "passive engagement" of users who keep in touch with each other while not actually communicating directly (see [74] for a more detailed discussion).

2.3.3 Social media sharing

As popularity of Social Network Sites grew, services focused on media sharing started including SNS features, turning into SNSs themselves. Examples are Flickr (photographs), Last.FM (music)[22], and YouTube (video)[23].

Stoeckl et al. [206] used an on-line questionnaire to analyze users personal motivation to share media content. According to their study, the most relevant motivation factors where "enjoyment, distribution of information, personal documentation and the desire for contacts"[24]. Van House studied the motivations for sharing media of Flickr users, basically via interviews[113], suggesting that the traditional social uses of personal photos were extended within this technology. Users still used the service for traditional "life chronicling", but also for new uses: maintaining social contacts (with close friends, mere acquaintances and even with strangers); self-representation; and self-expression.

Lange [138] investigated young people's video-sharing behaviour in YouTube. She gathered data from interviews, and performed a qualitative "ethnographic" analysis of posted videos, comments, and examination of subscription and contact practices. Lange observed that users manage their social network by their video-sharing practices, and how they enable other users to see their profiles by accepting friendship requests.

Agichtein et al. [7] aimed at quantitatively identifying high-quality content in social media, by using both interaction patterns and content-based lexical features.

[22] www.lastfm.com

[23] www.youtube.com

[24] This paper also provides a good review of previous studies on the motivations for user generated content.

2.3.4 Microblogging and Twitter

Microblogs are similar to blogs, but the content of an entry is typically smaller, allowing for quicker posting. A superbly successful microblogging service, Twitter, allowed users to do basically two things: post status messages shorter than 140 characters ("tweets"), and engage in a social network, in two ways, directly messaging another user, or subscribing ("following") to another user's posted messages.

Java et al. early analysis on the motivations for engaging on Twitter [116], follows a similar approach to what Broder did for the Web in [46], categorising user intentions in search queries. Java et al. classified Twitter users' intentions in: information sources (those users who provide information), information seekers, and friendship relationships. Java et al. analysis was quantitative, based on a crawl of the Twitter graph, and parsing the time-line of those users every 30 seconds to get their tweets, for 2 months[25]. Their method was to use the HITS algorithm [125] to identify hubs and authorities. Based on inspection of hub/authority scores of top-scored users, they proposed their rough categorisation of users' intentions. To study friendship relationships, they searched for overlapping communities in the undirected graph built from bidirectional links only. Using the Clique Percolation Method [179] (see section 2.5.2), they inspected the most frequent terms in the tweets of community members. They observe that "users participate in communities that share similar interests"; some community members would be providers of information while others would seek interesting information. By reading the posts, they manually categorize users intentions. The main would be: "daily chatter", conversations, sharing information or links, and reporting news.

Leskovec et al. influential paper [144] studied the dynamics of information propagation by tracking short, distinctive phrases that spread, in slightly modified versions, through on-line text[26]. Their approach is to build a *phrase graph*: each node is a phrase and links represent edits, from shorter phrases to longer phrases. They perform clustering to identify when a phrase is "the same", only edited. Then they track this phrase-variants-clusters in their dataset. These distinctive phrases happen to be quite abundant, and to show a significant diversity; the overall vocabulary, though, remains relatively stable. Leskovec et al. developed a mathematical model for the temporal variations the system exhibits, and found persistent temporal patterns in the news cycle: for instance, a lag of 2.5 hours in the peaks of attention to a phrase from news media to blogs.

Weng et al. [221] used the Latent Dirichlet Allocation model for topic distillation of the content of tweets, showing that the following relationship in Twitter is topically related, and thus, meaningful. The high reciprocity in Twitter is not caused just by "courtesy". Instead, there is *homophily*: users follow similar users. They also propose a ranking method of users according to their influence, by means of a "topic-specific" customized PageRank, with a weighting scheme that considers number of tweets published and topical similarity[27].

[25] It is a graph with 87 897 nodes and 829 053 directed links. They collected 1 348 543 posts from 76 177 authors who posted during the two months period.

[26] They monitored 1.6 million media sites and blogs for three months: 90 million articles overall.

[27] Their dataset is: 4 050 users who published more than 10 tweets, extracted from a set 6 748: the top

The largest-scale study of Twitter was recently made by Kwak et al. [134]. The authors crawled a huge part of Twitter (they argue it was the "entire Twitter site"): 42 million users as of July 2009. They showed that reciprocity in Twitter is low, indicating that what moves users is very different to friendship. They studied dynamically changing trends, and the spreading of a message broadcast through the network. A few users reach a large audience directly (*high degree hubs*); and the presence of these *hubs* is larger than what would be expected from a power-law distribution. Moreover, they suggest that users that do not have so many followers can still reach a large audience by the "word of mouth" mechanism of "retweeting" (repeating, maybe adding some comment, a tweet produced by another user).

It has been argued that Twitter is a valuable real-time source of news. In the context of earthquake detection, Sakaki et al. [195] have suggested that Twitter users might be perceived as *sensors*: aggregating information from user tweets, they show how earthquakes in Japan can be geographically located soon after they happen, and even propose doing so as an early detection public service. But it is a trustable source of information? Researchers Mendoza et al. [155] have also studied the tweets after an earthquake, in this case in the days following the disastrous 2010 earthquake in Chile. They study the propagation of false rumours through the network, by means of "retweets". The authors showed how the fact that a certain information is heavily "retweeted" does not make it credible. Not everything is lost, since they showed how false rumours where more questioned by other messages than rumours that were actually true. The latter where questioned initially by a small proportion of messages, and finally became supported by most of the messages.

It has been even argued that the aggregated knowledge contained in user tweets can be used to "predict the future". Asur and Huberman [23] suggest that social media content can be a sensor of the public opinion, and argue that their simple model, based on the rate at which tweets about a particular topic are created, can out-perform market-based predictors. To improve on this model, they suggest considering the positive or negative sentiment expressed in Twitter comments.

2.4 Social Network analysis

In this section, we review the history and current state of the art of network analysis. Networks are a powerful tool to analyze and explain complexity, and they do so by paying attention to what is *structural* in a system: the connections between individual components rather than the properties of components themselves. Good reviews of the field of (complex) networks are [167, 12].

We have decided to start from a simple definition of complexity (section 2.4.1), a physical sciences concept that underpins modern networks analysis. Then, we describe how network analysis has become an important tool to describe, understand and explain complex systems (section 2.4.2). Next, we introduce a series of examples of networks

1 000 Singapore-based users plus all their followers that are also from Singapore. They collect tweets for them: around 1 million.

from very different domains (section 2.2).

2.4.1 Complex systems

Recently, the idea of complexity has captured the imagination of researchers from different disciplines. But complexity is not itself a concept that allows for an easy definition, or one researchers from different fields agree on. Amaral and Ottino [17] propose a working definition of a complex system, as one that has the following properties:

1. Shows self-organization. It is composed of separate units, but they work together as a system.

2. Out of the interaction of the units comprising the system something *new* is created: this is the phenomenon of *emergence*. In words of P.W. Anderson “more is different” [20], “the whole becomes not only more but very different from the sum of its parts” [20]. Emergent behaviours can not be explained just by combination of the individual states[28].

This definition takes shape in contrast with other classes of systems. A *simple system* is one such that has a small number of components which act according to well understood laws. The pendulum is the epitome of a simple system, well understood and described by Newtonian physics. Nevertheless, simple systems are capable of generating complex dynamics. As an example, a forced pendulum, with gravity being periodic function of time, is a simple system that exhibits chaotic dynamics [17].

A “complicated system” is not to be mistaken with a complex system either. A system such as an airplane is complicated but not *complex*. A complicated system has a large number of components acting together, but these pieces have well defined, static roles, and their behaviour follows simple, well understood rules. A complex system, on the contrary, has a large number of components, which act according to rules that may change over time, and that might not be well understood. And the connections between components of the system can change over time, and the roles of elements might also be fluid. An example of complex system would be a flock of birds.

From an engineering point of view, a designer of a complex system, such as the airplane, ensures its ability to react to errors by building redundancy into the system. The system will be limited in its ability to respond to environmental changes. On the contrary, complex systems are capable of fluidly adapting to changes, and base on this their resilience to errors and adaptability to changes.

The idea of complexity is fundamental to understand the assumptions and methods of network analysis. Network scientists reject the “reductionist hypothesis” of the physical model of Newton and Laplace, according to which in order to predict the future state of a system it is sufficient to know the position and velocity of each of its particles. A clear flaw of the reductionist approach is one of scale: for a large system, such an approach

[28] Emergence is a concept that has been discussed in philosophy for a long time (see `plato.stanford.edu/entries/properties-emergent/`), but that has found its place in the physical sciences through the theory of complex systems.

would be computationally unfeasible. But a more fundamental critique is founded on the idea of complexity: the properties emerging from the whole system are "new", different from what can be explained by looking at individual components.

If the properties of the individual components are not enough to explain the behaviour of the system, where should we look at? The fundamental, almost axiomatic insight that network researchers from all fields share is that complexity has more to do with the nature of interactions than with the nature of the interacting objects themselves [201, p. 25]. For Social Scientists, social network analysis is based on the similar notion that the structure of social groups determines their behaviour, even more than the particular conditions of individuals (see also section 2.4.3).

2.4.2 Complex networks

Networks are a way of abstracting and representing the structure of interactions in a system, and have proved to be a very useful tool to describe, analyze and predict the behaviour of complex systems. In these section, we describe a few important phenomena of complex systems that have been studied in the literature by using network analysis.

Networks as a tool of explanation have a long history. Its first landmark is the foundation of graph theory by Leonard Euler, who in a paper published in 1736 solved the now famous Königsberg puzzle. The problem was set up on the Prussian city of Königsberg (now Kaliningrad, Russia) that has two islands on the Pregel River, connected to the main land by seven bridges. The question was to find a walk across the city that would cross each of the bridges just once. In order to mathematically prove that no such walk existed, Euler abstracted the structure of the problem, reducing the representation to the four land masses (the *nodes* of the graph, in modern terms) and the seven bridges joining them (the *edges*).

Euler laid in this way foundation for graph theory. Nevertheless, there is still a great difference between abstractions like Euler's and the complexity of real-world networks, like societies. The problem interested the mathematician Paul Erdős in the 1950s, leading to his fundamental contribution, in collaboration with Alfréd Rényi: *the random graph model*.

A random graph is generated by a random process. The random graph model is possibly the simplest useful model of real-world network [167] and had previously been studied by Solomonoff and Rapoport and Solomonoff[29] [202]. The Erdős-Rényi random graph model is an undirected network with n vertices, where the probability of two nodes being connected is p, independently at random. The probability of a pair of nodes not being connected is therefore $(1 - p)$. It has been shown that the distribution of node degree in the random graph model is a binomial, or a Poisson in the limit case for large n [167].

[29] Rapoport and Solomonoff proposed their "random net" a few years before Erdős and Rényi independently "discovered" the random graph, and gave it that name. Erdős and Rényi studied the model thoroughly and mathematically, and are usually credited with the discovery. Previously, de Sola Pool and Kochen had already studied mathematically random graphs, and described the "small-world" phenomenon, in a paper that circulated in the 1950s but was only published in 1979 [68].

The random graph model is useful in explaining, among others, two important characteristics of real-world networks: the *percolation threshold* and the *small-world phenomenon*, that we detail in the following.

Percolation Percolation is a concept borrowed from the physical sciences, where it refers to the filtering of fluids through a porous material. On the theory of networks, it refers to the question of under which conditions a graph shows connectedness, so paths from "top to bottom" of the network appear.

In this context the *percolation threshold* is the average number of edges per node necessary in order for a graph to be fully connected [201, p.32]. For the random-graph model, a suggestive phenomenon appears. The transition from a fragmented system to a connected one is brisk. If, in a network, edges are added one by one, there is a point when connectivity of the system suddenly emerges. This has resonance on the physical theory of critical phases: the transitions between a state of the system and a completely different one are sudden, and happen at critical points.

Percolation is very related to the phenomenon of resilience of networks: a network is resilient if it maintains its global structure when random edges are deleted from the network. For a more detailed discussion of these concepts, see [167].

The small-world phenomenon In the 1960s, Harvard's social scientist Stanley Milgram conducted his now famous "small-world" experiment. He wanted to answer the following question: if you choose two people randomly, how many acquaintances would be necessary to complete a chain from one to the other?

De Sola Pool and Kochen had already described the "small-world" phenomenon in their unpublished but heavily circulated paper, in the 1950s [68]. Based on records of 27 people, de Sola Pool and Kochen estimated the average number of acquaintances a person had was between 500 and 1500 people. Assuming an average of around 1000 acquaintances, they conjectured that any two randomly chosen US residents could be linked by "two or three intermediaries on the average, and almost with certainty by four"[68]. They were not able, though, to propose a way to test that suggestion.

Milgram chose three groups of so-called "starters" and gave them the task of sending a document to a "target" person, whom they did not know, by mailing it to someone they thought more likely to know the "target"[30]. Even though only 29% of the participant "starters" where able to actually deliver the message, the length of the chains started by those who did was surprisingly small: an average of five.

Since then, mathematical models of small-world networks have been proposed, and the concept has even reached popular culture, with the concept of "six degrees of separation" giving name to a theatre play by David Guare. And the small-world effect has been studied and verified directly in a large number of different networks.

[30]He chose as "starters" 100 stockholders from Nebraska, 96 randomly selected Nebraskans, and 100 people from the Boston area. Only 217 of them cooperated on the experiment. The "target" was a Bostonian stockbroker. The starters were given some information on the target, but they were only permitted to send the message to people they knew by name [91].

A way to evaluate the small-world effect on a network is to compute the mean distance (length of the shortest path) between all pairs of vertices. Let us consider $G = (V, E)$, where $V = \{v_1, v_2, \dots, v_n\}$ is an undirected graph with n vertices. If, for simplicity, we define the distance from a node to itself to be 0: $d(u, u) = 0$, the following formula gives the mean distance in a connected graph[31]:

$$l = \frac{2}{n(n+1)} \sum_{\substack{i=1,\dots,n \\ i \geq j}} d(v_i, v_j)$$

In his review of network theory, Newman lists up to 27 different networks from all domains and provides the value of l for them [167, see Table II, p. 10]. In many large networks, the value of l is small. Although the hypothesis of Milgram that *the world*[32] is indeed "small" seems very difficult to test empirically, after being tested on many different networks, it is assumed to be true by researchers. Maybe the largest small-world study is the analysis of the "who-talks-to-whom" network of Microsoft Instant Messenger users, performed by Jure Leskovec and Eric Horvitz while working for that company. Their dataset was composed of 240 million accounts on Microsoft IM[33], quite effectively all active users. They monitored the service for a month, and built a graph where nodes were users and an edge between two users represented that they had engaged in conversation during that period. The graph had a giant connected component, containing most of the users of the system, and the authors estimated, by sampling, that the average distance between pairs of nodes was 6.6.

The Erdős-Rényi random graph is the simplest model for the small-world phenomenon. Recall that in a random graph with n vertices, each pair of vertices is connected with probability p, and not connected with probability $(1 - p)$. For any node in the network, its average degree will be $z = p(n - 1)$. Therefore, any node, v, will have on average z adjacent nodes (neighbours at distance 1), z^2 neighbours at distance 2, and so on. The mean number of neighbours at distance d from the vertex is z^d, so if the network has n nodes, to go from the node to any other, a path of $z^d \simeq n$ steps is needed. So the distance through the whole network will be $d = \frac{\log n}{\log z}$, verifying the small-world property[34].

The small-world phenomenon can be more rigorously defined:

> "networks are said to show the small-world effect if the value of [the mean distance between pairs of nodes] l scales logarithmically or slower with network size for a fixed mean degree." [167, p. 11]

[31] If the network has more than one connected component, there are pairs of nodes that are not reachable from each other. A simple conventional way of still computing a value for l is to exclude from the sum pairs of nodes that are not connected by any path, although Newman suggests using the harmonic mean instead [167].

[32] In Milgram's research, the "world" was the US population.

[33] A very popular *on-line chat* service.

[34] The small-world effect is, we see, mathematically evident, for any network where the number of nodes within a distance d of a fixed node v grows exponentially on d, like in the random graph model. See also [36].

The Erdős-Rényi model is only the simplest to account for the small-world effect: it explains why distances in the graph are on average small on large graphs. The small-world model of Wats and Strogatz [219, 216] accounts for two seemingly contradictory features of real world networks:

- The number of steps between any pair of nodes in the network is small; this property is typical, as we have seen, of random networks.
- The number of triangles in the network is high. This property was first observed in social networks[35], and it is sometimes called *transitivity* (see section 2.1). This property is not present in random networks, rather, it is typical of ordered lattices[36].

Note that, in the definition of Watts and Strogatz, a small-world is a network that has *both* features. In the literature, it is called "the small-world model"; this model exhibits what we have called the "small-world effect" or "small-world phenomenon" *and* is highly transitive.

Watts and Strogatz measured the first property using the shortest distance between all pairs of nodes in the network and the second by the mean clustering coefficient of the nodes (see section 2.1 for definitions). They showed how real world networks had indeed short paths and high transitivity, and proposed a model for the emergence of these properties in simple networks.

Their model lays in between the completely ordered lattices and the disorganised random networks. Starting from a ring lattice with n vertices, connected to k edges each, they introduce randomness by a *rewiring* process. The rewiring procedure suggested consists of taking each edge in turn, and with probability p, connecting one of its nodes to a node chosen uniformly at random, except that no double-edges or loops are allowed. The rewiring process makes the graph "more random"[37] as p gets closer to 1. Watts and Strogatz showed that for a region of values of p, the small-world regime emerged: the number of steps needed to travel across the network from a node to another decayed rapidly, while the clustering coefficient was still high [219, 216]. A 10% rewiring made the distances in the network already much smaller [201, p. 40], keeping the transitivity high. In their paper, they tested the small-world effect in networks of very different nature: the electric power grid of Southern California and the neuronal network of the worm *Caeorhabditis elegans*[38].

[35] As early as 1953, Rapoport observed that if two people in a social network have a friend in common, there is an increased likelihood that they will become friends in the future (Anatole Rapoport. Spread of information through a population with socio-structural bias I: Assumption of transitivity. *Bulletin of Mathematical Biophysics*, 15(4):523–533, December 1953, cited by [74]).

[36] In a regular lattice the edges are distributed identically across the network. It is an apt model for regular systems like crystalline mollecular structures, but it is obviously very different from the structure of complex systems like societies. An example of one-dimensional lattice would be a network where every node has the same degree, so it is connected to the same number of neighbours.

[37] For $p = 1$, the graph is not actually a random graph (following the Erdős-Rényi model), although it is similar. The peculiar restrictions the model imposes, no loops and no double edges, makes it different from a random graph, and harder to analize mathematically. See [167, p. 28] for alternative models proposed in the literature.

[38] See section 2.2.

Barthélémy and Amaral [29] and Barrat and Weight [28] supported the ideas of Watts and Strogatz with analytical and numerical results.

Searchability The Watts-Strogatz model is a very valuable contribution, but it is flawed. One of its authors described it later as "unsuitable as a model of social networks" [217]. The first problem of the model was cleverly pointed out quite early by Kleinberg [126, 127]: the *searchability* or *navigability* of the network.

Kleinberg argues that, in the original Milgram experiment, it is not only striking that there is a path of short length that joins two "randomly" chosen people. It is even more surprising that the participants in the experiment were able to find it, having no global information (knowledge of the structure of the network as a whole), but only with local information (their circle of acquaintances). Just as in a random network short paths exist but it would be impossible to design a fast search algorithm capable of finding them using only local information, Kleinberg showed how no such algorithm existed either for the Watts-Strogatz model[39].

Kleinberg suggested a simple variation on the Watts-Strogatz model where fast decentralized search is possible. From a regular lattice (a square lattice in particular), shortcut edges are added between pairs of nodes. Rather than uniformly at random, as in the Watts-Strogatz model, shortcuts are more likely to join places that are closer in the Euclidean space defined by the lattice. The probability of a shortcut between two places in the lattice is proportional to r^{α}, where α is a constant and r is the distance between the nodes in the lattice (the L_1 norm, also called rectilinear distance, or Manhattan distance, over the lattice). Kleinberg obtained a lower bound on the average number of steps the greedy algorithm needs to find a randomly chosen target. The heuristic for the greedy algorithm is of course to chose, in each step, the connection that brings the message closest (in the lattice distance) to its target.

Kleinberg's model is hardly realistic for a social network, though. Other models have been suggested that allow fast decentralized search and are a better representation of social dynamics. Watts et al. [218] and, independently, Kleinberg [128], proposed a hierarchical "social distance" tree: people are grouped according to their "social characteristics", these groups are grouped into bigger groups, and so on. Social distance is then defined on the tree as how far up two people have a common ancestor[40]. Now, people have two kinds of information, the global social distance, and the paths to their local neighbourhood of acquaintances. The greedy heuristic would be to send the message to the neighbor that is closest in social distance to the target (most similar in the hierarchy of group membership).

[39] It is what Kleinberg calls decentralized search [126]. From an algorithmic perspective, the participants in Milgram's experiment used a "greedy" strategy to successfully find the target in a few steps, by sending, as instructed, the message to the person they thought most likely to know the target. In a random graph, it would be impossible to take this decisions. Of course, short paths exist, and a breadth-first-search algorithm would find them, but this would be equivalent to the participants sending the message to *all* their acquaintances, not just one of them.

[40] The "social distance" is not a distance in the mathematical sense (a metric), but rather a social notion individuals have of how far any other is from them.

The "reversal small-world" experiment, by Killworth and Bernard, supports that people navigate the social graph in this way, using social and geographical information [122]. Kleinberg showed that, for some parameter choices of the model, the search can be completed in $O(\log n)$ time [128], where $|V| = n$; and Watts et al. showed by computer simulation that a greedy algorithm is efficient over such a model for a range of parameters [218]. See also [167] for a more detailed discussion.

A particularly interesting experiment tries to tie the social experiments and the theoretical model for searchability that we have discussed. Adamic and Adar [4] simulated on real-world networks the search strategies users have reported using in small-world experiments, and correlated the success of the simulated strategies to the structure of the networks, and how well the models of Kleinberg and Watts et al. fitted them. They used two scenarios: a corporate email network and a student social network site. The greedy strategy (pass the message to the neighbour that is most similar to the target), succeeded in the email network. The authors attribute this partly to the correspondence of this network to the theoretical predictions of the models. In the on-line community, success was limited. This was in part due to the difficulty of finding similar users with the methods used by the authors, for this network.

Scale-free networks The second flaw of the Watts-Strogatz small-world model is that its degree distribution does not fit that of real networks. It has been shown that the degree distribution of many real world networks has a right-hand-side heavy tail: while most of the nodes have a small degree, below the mean, there is a long tail of nodes that have a degree far above the mean[41].

Power-law functions have been shown to fit the degree distribution of many networks. Let us formalize how this fit is made. A power-law is a discrete probability function, with probability mass function:

$$P(D = k) = k^{-\alpha}$$

for a constant exponent α, where D is the degree random variable. If we define p_k to be the fraction of nodes of the network with degree k, we say that the degree distribution of a network follows a power law if $p_k = k^{-\alpha}$. This means that the probability of a randomly chosen node of the network having degree k decays like a power of k, and the exponent gives the rate of decay. A smaller k means a slower decay, and therefore a longer tail. The cumulative distribution function of the degree, P_k, is defined as the probability that the degree is greater than or equal to k:

$$P(D \geq k) = P_k = \sum_{k' \geq k} p_{k'}$$

[41] For the random graph model of Erdös and Renyi, the probability of an edge between a pair of nodes is equal for all pairs. The degree distribution is binomial, and Poisson for the limit case of large n (so most vertices have the same degree) [36]. Ordered lattice networks have even more strongly peaked degree distributions. The model of Watts and Strogatz shows a peaked single scale degree distribution (see [167]).

If the degree distribution follows a power law, so does the cumulative distribution:

$$P_k \sim \sum_{k' \geq k} k'^{\alpha} \sim k^{-(\alpha-1)}$$

By plotting the cumulative distribution on a logarithmic scale, it is easy to experimentally see that a power-law fits the degree distribution: a perfect power law would be a straight line on that plot[42], with negative slope α.

The power law distribution is also known as Zipf's law after the linguist George Kingsley Zipf who first proposed it in the context of the word frequency distribution of a text corpus. It has since been studied for a broad range of applications (see [159]).

Barabasi and Albert[25] studied the actors collaboration graph, a crawl of the Web graph, and the US power grid, concluding that the three networks exhibited a power-law degree distribution[43]. They called this networks scale-free[44], and proposed a successful model of their generation.

Barabasi and Albert explained this phenomenon by the principle of *preferential attachment*, that can be intuitively reworded as "the rich gets richer": scale-free networks emerged in growing networks because nodes connect preferentially to the nodes that are already the most connected. The Barabasi-Albert model of the growth of a network is defined by this random process: nodes are added to the network one at a time and joined to a fixed number of existing nodes, where each node is chosen with probability proportional to its current degree in the network.

Scale-free networks are a subset of small-world networks. The distinguishing feature of scale-free is that the mean distance between nodes increases extremely slowly respect to the size of the network[25].

Scale free networks are highly robust, or *resilient*, in the sense that they are less affected by the removal of *edges*: Albert et al. [14] found that the ability of the nodes in scale-free networks to communicate is unaffected by high failure rates. Scale-free networks provide good communication and navigability, and do so due to the presence of highly connected nodes, or *hubs*. This good properties come at the cost of being very vulnerable to attacks: the removal of the highly connected *nodes* produces a great damage to the network.

Later work showed how certain factors might work against preferential attachment, making the network grow in different directions [18]: aging (for example, in the network of actors, they eventually stop playing), cost of adding links and limited capacity (for example, the network of airports), limits to information access (for instance in the Web, the interest on different areas might be skewed). Amaral et al. observed a cut-off in the power-law regime in real-world networks; the scale-free distribution may even

[42]For more details on power-law fitting see [54]. As we will see below, the long tail of some networks fits an exponential, rather than a power-law, distribution. They are similarly easy to spot, now on a semi-logarithmic scale. See [167] for more details.

[43]Amaral et. al showed that the tail of the US power grid network is actually exponential[18]. See note 45.

[44]The power law distribution is scale-free in the sense that a rescaling of the independent variable x changes the function by only a multiplicative factor: a function $f(x)$ is scale-free if $f(ax) = bf(x)$. See [167].

disappear. For instance, they presented empirical evidence that the electrical power grid was incorrectly classified by Barabasi and Albert [25] as a power-law network[45].

Barabasi and Albert work was so influential that a wealth of research emerged, and scientists were able to spot scale-free networks everywhere: the network of Web pages [13, 25]; the collaboration graph of movie actors or the electrical power-grid [25]; the "network of human language" [51], defined by Ferrer i Cancho and Solé as a graph were nodes are words and connections represent that the co-occurrence of two words is larger than expected by chance in a corpus of text; even code written in an Object Oriented programming language would produce a scale-free network, where nodes are objects and links are references among them [185]. For a more detailed discussion on the scale-free model see for instance [12].

Centrality Particularly within the context of social network analyis, the concept of centrality has been studied for a long time. In a social organization, the central members would be those able to influence a large number of people.

Botafogo et al. [42] studied hypertext systems previous to the most important of them, the World Wide Web. They defined notions of node importance based on their connectivity and on the centrality of the nodes in the hyperlink graph. From the definition of distance between two nodes in the network as the length of the shortest path, the radius associated to node v can be defined as the largest distance from v to any other node. Then, the centre of the network can be defined as the minimum radius node.

This concept has been applied to the academic citation network to estimate the most important articles as those with a small radius (see [55] for a more detailed discussion).

In the context of the Web, considered as a directed network, Kleinberg defined index-nodes as those with a high out-degree (over the network average); and reference-nodes as those with high in-degree [125]. A difficulty with this definitions is what to do if the network is not connected, and therefore, there are pairs of nodes that cannot be reached from each other. Botafogo et al. [42] suggest substituting those infinite distances with a constant value, K. In the modified node-to-node distance matrix they define out-centrality of node v_i as the sum of the distances on the i-th row of the matrix, and a node is central if this measure is relatively small.

For a detailed discussion on centrality indices and algorithms, see [45].

2.4.3 Analysis methodology

Network analysis is an interdisciplinary field. Networks have been studied for a long time in the social sciences, and mathematical formalizations have been proposed since the 1950s. And then, particularly in the last two decades, other disciplines have also focused on them: computer science and mathematics; the physical sciences (statistical physics in particular); biology; and even fields like business management, or criminology.

[45] They classified small-world networks in three sub-classes [18]: (1) scale-free networks, characterised by power-law tail; (2) broad-scale or truncated scale-free networks, where the tail has a cut-off; and (3) single-scale networks, with a fast decaying tail, such as exponential or Gaussian. The Californian electrical power grid would be in the third group, with an exponential decay.

Social networks, those that represent the structural connections of people, have been studied by scholars in all these fields.

Despite this wealth of research, the communication between different fields has been limited. The different goals, approaches and methods used by researchers from different traditions have kept the lines of research somewhat separate. It seems that they could obtain important insights from each other.

In his book on the history of network analysis in the social sciences, Freeman uses the methodology of network analysis itself to verify the intuition that research in this topic from physicists and social scientists has been historically separate [88], in a small-case yet interesting experiment. He mapped the citation network from a collection of papers dealing with the small-world phenomenon. He removed those nodes (papers) that were isolated (neither cited nor were cited by any other in the collection). Freeman divided the remaining 395 papers in the collection in two groups, depending on whether the writers belonged to the "physical sciences" or the "social sciences" traditions, and studied the citation patterns. It turns out they conformed two clearly defined communities. The citations in papers from each community stays within the community about 98% of the time.

Freeman argues that this separation leads to wasted effort (reinvention and rediscovery). He gives as examples two important papers: Watts and Strogatz paper on the small-world phenomenon cites Milgram's work, but remains oblivious to the first formulation of the phenomenon by de Sola Pool and Kochen[68], and the many follow-up papers on the subject. The observation of the skewed degree distribution of real world networks by Barabasi and Albert in their relevant paper on scale-free networks would have been described already by Paul Lazarsfeld in 1938 (Moreno, Jacob L., and Helen H. Jennings. 1938. "Statistics of social configurations." Sociometry 1:342–374., cited in [88]); and their model of preferential attachment would be indebted to work by Derek de Solla Price as early as 1976 [88]. On the positive side, these two papers had an immense impact on the physics community, and succeeded in interesting physicists on the structural approach to social phenomena [88]. And the split seems to be reducing, as researchers in other fields seem to be now more aware of the social network tradition [88]. Physicists, like Newman, acknowledge that "of the academic disciplines the social sciences have the longest history of the substantial quantitative study of real-world networks" [167].

Theory of networks has been a fruitful area of research for the social scientists, and goes back quite a long time. We briefly review now how social scientists have developed explanations to social phenomena based on network interactions. From that historical landscape, the intuitions, assumptions and methodological approach of the field arise. We will make them explicit at the end of the section. For a comprehensive historical review of Social Sciences analysis of networks, see the work by Freeman [88]. For a discussion on social network analysis concepts and methods, see [215].

We agree with Borgatti et al. [40] in that the power network theory has brought to the social sciences comes with the idea that the fabric of social interactions between individuals explains some interesting social phenomena:

> "Social network theory provides an answer to a question that has preoccu-

> pied social philosophy [...] the problem of social order: how autonomous individuals can combine to create enduring, functioning societies." [40]

In his report on the history of network analysis in the social sciences, Freeman [88] accounts for very early intuitions into the importance of the structure in society: the French social philosopher who in the view of Freeman was the first scholar who looked at society in terms of the interconnections among social actors. A more explicitly structural view on society organization appears in the work by Georg Simmel:

> "If, therefore, there is to be a science whose subject matter is society and nothing else, it must exclusively investigate these interactions, these kinds and forms of sociation". Simmel, Georg. 1908/1971. *On Individuality and Social Forms*. Chicago: University of Chicago. pp. 24–25 cited by Freeman in [88].

An example usually pointed out (see [40, 192]) as an early example of the use of networks to explain social phenomena is the work of Jacob Moreno and Helen Jennings, who in the 1920s 1930s studied the social networks of the children at the Hudson School for Girls in upstate New York [163]. They used what Moreno called "sociometry" to graphically represent the feelings expressed by individuals to each other. This representation, or "sociogram", let them observe the different roles of individuals in the network of their social connections[46].

In the late 1920s, a group of researchers around Harvard university made considerable contributions to the quantitative analysis of social networks. W. Lloyd Warner studied the interpersonal network of the industrial city of Newburyport, Massachusetts [88]. Warner later influenced the research of Elton Mayo at the company Western Electric; while Mayo's research on industrial productivity followed initially a purely psychological approach, focusing on the individuals, Warner suggested to account also for the social structure, the patterns of interaction between workers [88]. Years later, another Harvard scholar, George Caspar Homans, had important insights in the sociology of interaction and sentiments. In his book, The Human Group [112], Homans studied the structure of social groups and the positions of individuals within those groups (an early study of social communities; see section 2.5).

Mathematical breakthroughs in the analysis of social networks came in the 1940s and 1950s. Alex Bavelas created at MIT the Group Networks Laboratory; their work displayed back then all the traits that are found in contemporary social network analysis: the structural hypothesis; the collection of empirical data on social interactions; and the

[46]The Hudson School for Girls hosted between 500 and 600 girls sent there by the courts of the state of New York. Moreno and Jennings recorded the liking or disliking feelings from the girls to each other, together with spatial information (the "cottages" where the girls lived together in groups). They used the resulting diagram to explain why a unexpectedly high number of girls decided to run away from the institution, arguing that it was their position in the network of social interactions that determined their behaviour.

While some where central to the network, others remained peripheral and isolated. This concept was adapted later into the idea of "network centrality", that is pervasive in the social sciences study of networks.

use of graphs to represent the patterns of interaction. The Bavelas group developed a formal model for centrality.

But up to that point, the social network perspective had not reached the social science community. It was on the 1940s through the 1960s, that the network approach was embodied, due to the work of a series of researchers [88].

In the 1950s, Kochen, a mathematician, and de Sola Pool, a political scientist, wrote their heavily circulated paper, unpublished at the time, describing the "small-world" problem [68]. Stanley Milgram's experiment tested their suggestion, as we have seen (section 2.4.2).

Another fundamental idea coming from the social sciences is the "weak ties" theory of Granovetter [101]. Studying interpersonal connections, Granovetter suggested that in maintaining the social cohesion, not only the strong ties matter. On the contrary, the weak ties, those between acquaintances rather than close friends or family, are more efficient to convey novel information. Granovetter did not provide empirical proof or mathematical model to support his conjecture. But his ideas were more recently popularized as a theory of "social capital": capitalist competition and entrepreneurship [50].

In the 1990s, the field of network analysis was finally well established within the larger discipline of Social Sciences, with conferences, journals and an American professional organization [88].

Let us now summarise what characterizes network analysis from the social sciences:

- Assumptions:
  - Structure determines behaviour. The underlying assumption (taken almost as an axiom) of network analysis from the social sciences is that to understand the behaviour of people and social groups, the researcher has to look at the structure of the social group: the network of connections and interactions. The particular, psychological, conditions of the individuals studied by traditional social scientists are not enough to explain certain social phenomena.
  - The ties in the network act as communication channels: they allow for the node-to-node transmission of information, knowledge, diseases, etc.
- Research questions:
  - Social scientists expect the structure of social networks to be different from a context to another. The properties studied are expected to be different if the networks are qualitatively different. The social scientist will try to explain why this is so.
  - The concept that has received most attention is "centrality": identifying those nodes that have a crucial role in the social network, caused by their position and connections.
  - There is a fundamental interest in the dynamics of networks: how they grow, how ties are formed, how they gain or lose strength.

  - As we mentioned in 2.2, social scientists have distinguished a great variety of ties in social networks. Furthermore, the same kind of tie can be of varying strength.

- Approach: social scientists study small networks, but in great detail. Surveys and interviews are the preferred methods of data gathering. Early research was typically descriptive, and contained no quantitative analysis or mathematical theory to support it. Later, social scientists have incorporated mathematics and statistics, and have proposed a great variety of measures of network features.

A common critique to the social sciences approach to network analysis is the difficulty of obtaining reliable data. Costly methods like surveying and interviewing limit the size of the available data. Moreover, it is argued that methods like interviewing suffer from biases, as the answers are subjective [167]. So research, and particularly research coming from other disciplines, like computer science and the physical sciences, has turned to different methods of data gathering. First, collaboration graphs were studied, like the paper co-authoring or the actor collaboration. Paper co-citation was also studied. Then, people communications were a source of large datasets: telephone calls, email, instant messaging. At last, the appearance of community-based web services and Social Network Sites (see section 2.3), turned out to be a source of large scale datasets at a very low cost.

Physical scientists have taken a very different approach to the study of social networks, that goes beyond the datasets used. To understand it, it is important to situate the study of social networks in the context of the study of complexity (see section 2.4.1). The particular approach of statistical physicists to complexity is responsible of their take on network theory.

One of the guiding principles of statistical physics is the concept of "universality". For statistical physicists, it is possible to classify all critical system into "universality" classes. They try to find properties that are universal, that is, common to complex systems irrespectively of their nature (see [17]). So research conducted by physicists on networks commonly consisted on showing how networks from totally different domains (social, natural, technological) have a certain property, that is not present in random graphs [40].

On the other hand, network physicists share the assumptions of social scientists: the structural hypothesis and the assumption of links as communication channels.

2.4.4 Availability of data

Availability of data for the research community is crucial to obtain and test significative results. The quality of the data available limits the validity of the experimental results obtained. As we have discussed, the traditional social sciences approach to the study of networks relied on labor-intensive gathering of data through surveys and interviews. The rise of social computing, with millions of users interacting with each other in Social Network Sites, media sharing services and so forth, promises the availability of large-scale datasets. But the distribution and availability of social network datasets has severe difficulties: privacy concerns, difficulties in gathering representative samples and increased distribution costs [196].

On August 2006, a well intentioned but ultimately ill fated research project showed clearly the dangers of releasing user sensitive data for research purposes. Dr. Abdur Chowdhury, head of AOL Research, decided to release part of AOL query log. The data contained the queries issued by over 650 000 users over a period of three months. Before the company reconsidered this decision a few days later, and removed access to the data, it had been already distributed over the Internet, and it is still available in mirror sites.

The AOL researchers had been careful to remove from the data set the identities of the particular users, and substituted them with supposedly anonymous numerical identifiers. But as the queries were linked to a particular numerical identifier, it was possible to reconstruct, from the list of issued queries, the identities of the users. Reporters of The New York Times were able to do exactly that [27]. A class action file-suit was filed against the company, seeking a massive compensation for the users [173]. And researchers were reluctant to use the dataset, given the privacy of the users had been violated by distributing it [109].

Since then, there has been significant research on anonymizing social data. The aim is to remove any user identifiable information, but keeping enough of its structure so research on the anonymized set is still significant. The problem is not solved: it has been shown how some of these techniques can be reverted, and user sensitive data could be exposed even in thoroughly anonymized sets [24, 164, 165].

As a consequence, practically no datasets have been distributed since then.

Privacy of the users is not only a legal or business concern. Over the damage to a company that a disclosure of private data would cause, ethical issues are of great importance. Users entrust their data to companies offering a variety of services, and they have to be protected from violations to their privacy[47].

As datasets are owned by private companies, researchers in the academia do not have access to multiple datasets to test their results for statistical significance. Researchers working for these companies have access to the data, but they are not able to distribute it, so the verifiability of their results cannot be tested thoroughly.

A solution academic researchers have been using is to crawl themselves the data publicly exposed in Social Network Sites. This method is not completely reliable, as it is difficult to obtain a representative sample, and companies move towards disallowing automated crawlers (introducing technical limitations and explicitly forbidding it through the terms of service of their APIs). The same privacy concerns that affect company owned datasets are valid for crawled datasets. The fact that the crawled data was pub-

[47] It is interesting to observe that, when engaging on a SNS, for instance, a user is giving data of different types. Bruce Schneier's taxonomy distinguishes[199]: (1) service data, given to the site in order to register; (2) disclosed data: what the user shares or posts on his own profile; (3) entrusted data: posted to other's profiles or pages, so the user creates it but loses control of it; (4) incidental data: is created by others about the user; (5) behavioural data, is the data of the user activities collected by the site; and (5) derived data, which is derived from the user's behavioural data. Schneier defends that different data requires different protection and policies.

licly accessible when the crawled was performed do not free the researcher from taking care of the privacy of the users.

Another difficulty that cannot be overlooked is the technical difficulty and cost of sharing multi-gigabyte datasets.

A promising alternative are synthetic datasets. The idea is to generate, from a real social graph, and an appropriate model, a completely randomized synthetic graph that has the statistical properties of the real graph. The graph is matched by feeding the real graph features as parameters to the model. Sala et al. have designed models to generate such graphs from social network datasets, and showed ways to evaluate the fitting of the synthetic graph to the original [196].

This idea is similar to the use of *benchmark graphs* (see section 2.5.4). The approach is promising, but it has to be tested empirical that is indeed applicable: if a relevant, but not yet described, feature of the real social network is not kept by the model, it would be impossible in practice to discover it from analyzing the synthetic graph.

2.5 Communities in Networks

Real world networks are different from the simplified models we have discussed in section 2.4.2. They are typically inhomogeneous, the distribution of edges is not only skewed globally, but also locally: there are regions of the network where the density of edges is high, and these groups have a low concentration of edges between each other. This feature is what we call community structure [84]. Communities are, intuitively, dense regions in the network that form groups, with a lower density of edges between different groups [215].

In the previous sections of this chapter, we have described the general background of network theory, with a particular interest in social network analysis. In this section, we describe the state of the art of community detection, as one of the main goals of our investigation is to study the community structure of endorsement social networks. As we will see, there is a very large body of research on the subject, but the problem is not yet solved. Quoting Fortunato and Castellano, the ideal method would be one that [84]: "delivers meaningful partitions, and handles overlapping communities and hierarchy, possibly in a short time". But "No such methods exist yet". In case the network is directed, regarding directionality naturally is another good property of the desired method, and also a much less studied problem.

As we will see, our initial goals (see section 1.4) are open questions: dealing with very large networks, considering directionality and overlapping communities. In this section, we will define the family of community finding problems (section 2.5.1), and review the state of the art of community detection methods (section 2.5.2). We will discuss the computational complexity and performance of the methods, and special sections will be devoted to those particular problems that are our research goals: directed networks and overlapping communities. Then, in section 2.5.3, we discuss the second open problem we have studied: measures of influence, and different roles in social networks, in relation to their position within communities.

Community detection has multiple applications in every domain of network analysis: sociology, physical sciences, biology and computer science. The difficulty of the task is not only algorithmic, but also conceptual: the concept of community is, as we shall see, ambiguous, and no unique definition is shared in the literature.

Analyzing community structure is one of the most active network research areas. It has a long history in the social and physical sciences, and has experimented an "exponential growth" recently[48]. So it would be unfeasible to review all relevant work on the topic. Our aim is to describe and exemplify some of the most important approaches that have been proposed. For good reviews on the subject, see [168, 66, 198, 184, 84].

Fortunato [84] traces the analysis of community structure back to 1927, when Stuart Rice [191] searched for "blocs" or "groupings" within small political bodies based on the agreement of their voting patterns. The method was rudimentary, as it implied computing the degree of agreement between all pairs of members, removing those combinations that would not yield a large agreement value, compute the agreement for larger groups and so on[49]. The first definition of *clique* (see 2.1) comes from the social sciences: Luce and Perry formally defined in 1949 the notion of clique as a maximally complete subgraph of a network [151].

In 1950, Homans [112, p. 83] devised a method for community detection: reshuffling rows and columns in a matrix representation of the recorded interactions among a number of people. This is a clear antecedent of the now standard method of finding blocks in a matrix [168](see the spectral bisection approach, in section 2.5.2).

Later, in 1955, Weiss and Jacobson [220], starting from a matrix representation of the workplace relationships among the employees of a government agency, tried to find what they called "work groups". A work group was defined as a "set of individuals whose relationships were with each other and not with members of other work groups except for contacts with liaison persons or between groups"[220]. Their approach was to remove from the relationships matrix this liaison persons, that act as bridges between groups, isolating the work groups. This idea of removing nodes that operate as connectors between communities is, as we will see, also used in a number of modern community detection methods. One of them was described in a seminal paper [98], with which Girvan and Newman put the issue of community finding at the forefront of network research, and proposed a successful algorithm (see section 2.5.2). The paper triggered the interest in the field, yielding in the last decade a wealth of research, and a huge variety of community finding methods. The availability of large scale data, with Social Network Sites (section 2.3), where communities appear naturally, has also driven a lot of the recent interest.

[48]Porter et al. describe the production about this particular problem has been so active over the last decade that "new discussions or algorithms [were] posted on the arXiv preprint server almost every day." [184]

[49]Rice analyzed small sized structures, like New Jersey Senate, composed of only 21 members [191].

2.5.1 Definitions of community

There is not a unique definition of community largely shared by researchers. Rather, there are many different ones. Following Fortunato [84], we can roughly classify definitions in three groups: local definitions; global definitions; and definitions based on node-to-node similarity:

- We call local definitions those where the community is a subgraph defined only by looking at the properties of its nodes and, possibly, its immediate neighbourhood. These definitions are typical of social network analysis.
  - Self referring definitions consider the subgraph alone: a community is then defined as a maximal and highly connected subgraph.
    A ***clique*** is a maximal connected subgraph (see 2.1). Triangles are the simplest cliques, and they are very frequent, particularly in social networks. Larger cliques are less common, and a very simple way of defining a community would be a large clique. Finding cliques in a graph is an NP-complete problem [84]. This definition may be too restrictive: a subgraph that is "almost" a large clique, except a few edges are missing, may be still an interesting structure. The following definitions relax the completeness condition of a clique: an ***n-clique*** is a maximal subgraph such that the distance between any pair of vertices is at most n; a ***k-plex*** is a maximal subgraph such that each node is adjacent to all others, except at most k; a ***k-core*** is a maximal subgraph where each node is adjacent to at least k other.
  - Comparative definitions: the connectivity of the nodes in the subgraph candidate to be a community is compared with the connectivity to nodes not in the subgraph.
    For instance, Radicchi et al. [187] defined ***strong community*** as a subgraph where *each node* has more neighbours inside than outside the subgraph[50]; and ***weak community***, as a subgraph where the total degree of the nodes inside is larger than the total external degree (the number of edges to nodes external to the subgraph).
- In global definitions, the structure of the entire graph is considered:
  - An important class of global definitions are ***null models***. A null model of a network, in this context, is a graph that shares some features with the original, but has no community structure. Communities are defined, then, as structures present in the network but not in its null model. The most widely known null model supports the definition of Newman and Girvan ***modularity*** [171]: it is obtained by randomly rewiring the edges of the original network, but keeping the degree of each node. A community is then defined as a subgraph that has more internal edges than a subgraph with the same nodes and respective degrees would have in the null model.

[50] This definition was introduced by Flake et al. in their study of Web communities [83].

The same null model is used by Arenas et al. [22] to generalize the definition of modularity, by counting not just the number of edges internal to the community, but the number of small distinctive subgraphs (called *motifs* [158]). Reichardt and Bornholdt [189] have designed a general framework of null models, that includes but is not restricted to modularity.

- Definitions based on node-to-node similarity: communities are groups of vertices which are similar to each other according to a quantitative criterion. One can compute the similarity between each pair of vertices with respect to some reference property, local or global, no matter whether they are connected by an edge or not. Each node ends up in the cluster which nodes are most similar. Similarity measures are at the basis of many traditional methods, like hierarchical, partitional and spectral clustering. A few options are:
  - Spatial metrics work by embedding the graph in a metric space. Then, any L_n norm can be used, like the Euclidean L_2, the Manhattan distance L_1 or the L_∞ norm. Another frequently used similarity is the cosine similarity.
  - Alternatively to embedding the graph in space, a similarity can be defined based on the adjacency relationships between nodes. A possible definition is based on the principle of structural equivalence. Two nodes are said to be structurally equivalent if they have the same neighbours. Nodes that have large degree but do not share neighbours are considered very distant. Also related to structural equivalence is the Pearson coefficient of the rows, or columns, of the adjacency matrix.
  - Another measure of similarity is the number of edge-independent paths between two nodes. Two paths are edge-independent if they do not share any node. Analogously, the number of node-independent paths can be used. These similarity measures are related to the maximum flow/minimum cut theorem.

Finally, there are many community detection algorithms where communities are not defined in principle: they are what results from a defined detection procedure. This is the case of divisive algorithms.

2.5.2 Community detection

In this section, we classify a number of approaches that we consider remarkable. For more information on these and other methods, see the review articles: [66, 198, 184, 54, 84], among others.

Most techniques are designed for undirected, unweighted graphs. We first review these approaches, and later we deal with methods that tackle directed graphs.

Traditional clustering techniques

The problem of graph clustering, or graph partitioning, is rather a family of similar problems. There have been a great number of clustering techniques proposed to solve these problems. We review first a few traditional graph clustering approaches.

Data clustering A possible definition of the problem would be: try to maximize the similarity of the nodes within the same subset, while at the same time minimizing the similarity to those in others. This task is similar to data clustering, a problem that has been studied for a long time from the statistics and machine learning communities. In data clustering, one tries to organize the data into groups of similar objects.

A popular method of data clustering is *k-means*. This is a partitioning algorithm, where the aim is to divide the data space in a fixed number of groups, k, so each data point belongs to the nearest group, where the location of a group is defined as its average position (centroid). The parameter k has to be given as an input. Then, k centroids are defined, say randomly. Each centroid represents a cluster. Then, on each iteration of the algorithm, data points are assigned to the nearest cluster (centroid). The centroid of each cluster is recomputed as the average of the data points in the cluster, and then the process is repeated, until some stopping criterion is met (for instance, the clusters stabilize, and there are no examples moving from a group to another).

On a network, provided that a node-to-node similarity is defined, this method can be applied: nodes are embedded into a metric space, so that each node is a data point, and there is a distance (or a similarity) defined on the pairs of nodes [84]. A natural choice for a similarity in a network is the (normalized) weight of the edges.

A drawback of methods like k-means clustering is that it is necessary to specify as an input parameter the desired number of clusters in the solution. On the other hand, defining a similarity between the nodes might be natural for some networks but not for others.

Graph partitioning Another formulation of the graph clustering problem consists of dividing the set of nodes of a graph into k disjoint subsets, in a way that minimizes the number of edges between the subsets. Partitional techniques try to find the clustering solution by successively partitioning the graph in a number of subsets, and applying recursively the method to the resulting clusters. Pothen [186] reviewed these methods. Pothen observed that, while "most variants of the graph partitioning problem are NP-hard"[186], some of the approximate algorithms give good solutions, although not necessarily optimal.

Again, a major drawback of these methods is that it is necessary to give as an input the desired number of partitions.

One of the earliest graph partitioning methods is the Kernighan-Lin algorithm [121]. Kernighan and Lin defined the problem of partitioning the set of nodes of a graph with costs on its edges into k subsets, minimizing the sum of costs on all edges cut. Kernighan-Lin algorithm obtains a partition of a graph in two subsets. More partitions can be obtained by recursively applying the bisectioning method to each of the subsets.

For simplicity, let us consider that all edges have equal costs. The algorithm starts with a partition in two sets, that can be random or informed by some initial criterion. Then, at each iteration, the algorithm swaps subsets of the same number of nodes from a set to the other, in a way that reduces the number of links between both sets. The algorithm stops when there is no possible gain in swapping node subsets, or a maximum number of swaps has been reached. If we assume a constant number of subset swaps at each iteration, the Kernighan-Lin bipartition algorithm requires $O(|V|^2 log\,|V|)$ time [121]. On sparse graphs, a modified heuristic runs in $O(|V|^2)$ time. The algorithm is heavily dependent on the initial configuration, so it is customarily used to *improve* partitions obtained with other methods.

Fiduccia and Mattheyses designed an efficient implementation to compute the bipartition of the Kernighan-Lin algorithm in $O(|E|)$ time, by using doubly-linked lists as data structures [81].

METIS[51] is a popular software solution for graph partitioning. It tries to find a cut of the minimum number of edges to obtain two disconnected components of similar size (also called a *minimum conductance cut*) [120].

A well known result of graph theory is the max-flow/min-cut theorem. Let us think of a network as a pipe system, where there is certain commodity flowing through the edges, from a node, called source, to another, called sink, and each edge has a fixed capacity. The max-flow/min-cut theorem states that the maximum amount of flow passing from the source to the sink *equals* the minimum capacity that needs to be removed from the network so that no flow can pass from source to sink. The maximum capacity, or minimum cut, is the sum of the capacities of the removed edges. Ford and Fulkerson proved in 1956 this theorem, and proposed an algorithm to compute the maximum flow[52], commonly known as the Ford-Fulkerson algorithm. The idea of the algorithm is simple: while there is a path from source to sink, with remaining capacity on each of its edges, we send as much flow as possible, and subtract that amount from the available capacity of the edges in the path. If the capacities are integers, the complexity of the algorithm is $O(M \cdot |E|)$, where M is the maximum flow, and E is the edge set[53].

Flake et al. [82] used a max-flow/min-cut approach to find communities. They defined a community to be a subgraph where each of its nodes has more edges within the subgraph (inter-community edges) than to the outside (intra-community edges). This is the *strong community* definition of section 2.5.1. The method requires to know in advance a few nodes that belong to the community (*seeds*). Then, they find the minimum cut that disconnects the graph, so the known nodes are in the same connected component. From this component, a new set of seeds is chosen, and the procedure is repeated. How to choose the seed nodes? Flake et al. use the HITS algorithm [125], and the linked hubs and authorities that it yields are taken as seeds of the communities. Flake et al. applied their method to the problem of finding communities in the Web (see section 2.5.2).

The max-flow/min-cut theorem can also be used to find a minimal cost partition in two sets of unspecified size, that is a lower bound for other partitioning solutions.

[51] Visit glaros.dtc.umn.edu/gkhome/views/metis

[52] P. Elias, A. Feinstein, and C.E. Shannon also proved the theorem, independently, in the same year.

[53] A path from source to sink can be found in $O(|E|)$ time, by breadth-first-search, for instance, and this has to be done (with integer capacities) M times as much, because each time, at least 1 unit of flow is sent through the found path.

Hierarchical clustering It is difficult to decide in advance how many clusters would be desirable. A possible workaround is to make additional assumptions, like using some external information to decide the number of clusters, or impose constraints on the size of the resulting clusters. These assumptions might not be clearly justified. It is possible to imagine that the community structure of a network is not as simple as a partition into equally sized groups. It has been shown [197] that some networks have a hierarchical structure, so small communities merge together to form large clusters.

Hierarchical clustering algorithms allow for uncovering this multi-level community structure, and do not require the number of communities as an input. These methods need a node-to-node similarity defined on the graph. Computing the similarity between all pairs of nodes, a similarity matrix is built, and the methods work on this matrix.

Hierarchical clustering algorithms can be further subdivided in agglomerative or divisive. Agglomerative methods are "bottom-up": they iteratively merge clusters that are similar into larger clusters, until a unique cluster with all nodes is made. Divisive algorithms work in the opposite direction, "top-down": clusters are split into smaller ones. Agglomerative methods are usually preferred in this context.

There are different options to compute the similarity between a pair of clusters. The common idea is to use the similarities between pairs of nodes where each of the nodes is in one of the clusters to compare. *Single-linkage* defines the similarity between two clusters as the minimum similarity of any such pair of nodes. *Complete-linkage* uses the maximum pair similarity instead. Finally, *average-linkage* defines cluster similarity as the average similarity between pairs of nodes.

The solution of a hierarchical clustering algorithm can be represented as a tree, called *dendrogram*, where each node is a cluster. A hierarchical divisive algorithm starts from the root of the tree, the entire graph, and divides it in smaller clusters. The child relationship of the dendrogram represents this successive partition. In a hierarchical agglomerative algorithm, the process is usually started from the individual nodes, that are the leaves of the dendrogram. Clusters are merged into bigger ones, and the parent relationship in the dendrogram goes from the smaller clusters to the larger cluster formed by successively merging them.

The software toolkit CLUTO implements hierarchical clustering algorithms [119].

Divisive algorithms

Divisive algorithms work by removing inter-community edges until the communities in the network are disconnected. Depending on how the inter-community links are detected, different algorithms have been proposed in the literature. This is basically the same idea of divisive hierarchical clustering. The main difference is that, in hierarchical clustering, nodes that are dissimilar according to some measure get disconnected when a cluster is divided in smaller ones. Here, edges that match some property are removed.

Algorithm of Girvan and Newman The best known divisive algorithm was proposed by Girvan and Newman [98]. Inter-community edges are detected according to the value of their *edge betweenness.* Edge betweenness of an edge is the number of shortest paths between node pairs in the network that contain the edge. This measure extends to edges the concept of point (node) betweenness proposed by Freeman in 1977 [87].

The algorithm, like other divisive algorithms proposed later, works iteratively by: (1) computing the edge betweenness for all edges; (2) removing the edge with the highest value; (3) recompute all values of edge betweenness for the new graph; and (4) iterate.

Since computing the shortest path between a pair of nodes takes $O(|E|)$ time using breadth-first search, and there are $O(|V|^2)$ node pairs, calculating edge betweenness for all edges would take $O(|E||V|^2)$ time. But there are algorithms that can compute all edges betweennesses in $O(|E||V|)$ time [166, 43, 231].

This computation has to be repeated once for each edge removed from the network, so the running time of the algorithm is worst-case $O(|E|^2|V|)$. On a sparse graph, complexity will be $O(|V|^3)$. In practical terms, the algorithm is usable for networks up to about 10 000 nodes [167].

Girvan and Newman proposed in their paper two other measures of betweenness. The first is current-flow betweenness, defined by considering the network as a resistor network, where edges are resistances. If a voltage is applied between a pair of nodes, each edge carries a certain amount of current, that can be computed by solving Kirchoff's equations. Repeating the computation for all pairs of nodes, the current-flow betweenness is the average current carried by an edge. In this case the complete community detection process takes $O((|E|+|V|)|E||V|^2)$ time, or $O(|V|^4)$ on a sparse graph [167]. The second is a random-walk model: betweenness is defined as how frequently a random walker on the network would follow each edge. The complete calculation takes now $O(|V|^3)$ on a sparse graph, and they showed that the calculation is equivalent to the current-flow betweenness [167].

The original version of Girvan-Newman's algorithm yielded a hierarchy of partitions. The authors later proposed choosing as solution the partition with the highest modularity (see section 2.5.2); this criterion has been frequently used in the applications of the algorithm.

Tyler et al. [212] modified the Girvan-Newman algorithm, improving on computational complexity at the cost of introducing randomness. Instead of computing edge betweenness from the paths between all pairs of nodes, they used a random sample of the nodes. The method, after aggregating over multiple runs, showed good approximations in empirical studies. They tested their approach on an email network of nearly 1 million nodes.

Information centrality Fortunato et al. [86] proposed an alternative measure of the importance of the edges: information centrality. Their algorithm is based on the concept of efficiency of information transportation in networks, as defined by Latora et al. in [140]. It is assumed that information travels along the shortest paths of the network, and the efficiency of nodes u and v is defined as the inverse of their distance (length of the shortest path from u to v). The efficiency of the graph is then defined as the average of the efficiencies of all pairs of nodes in the network. Information centrality of an edge in the network is defined as the decrease in efficiency produced by removing the edge from the graph.

In the algorithm of Fortunato et al. [86] edges are removed, one by one, in decreasing order of information centrality. The method is analogous to Girvan-Newman, but slower: $O(|V|^4)$ on a sparse graph. The authors argue, though, that it gives better solutions when the communities are heavily interconnected.

Edge clustering coefficient Radicchi et al. [187] modified the Girvan-Newman algorithm, using the edge clustering coefficient to choose the edges to be removed, which is an analogue of the clustering coefficient for nodes defined by Watts and Strogatz [219] for the case of edges. The *edge clustering coefficient* is the number of triangles to which an edge belongs, divided by the number of triangles that it could belong to, taking into account the degrees of its nodes. Intuitively, inter-community edges have smaller values of edge clustering.

This version of the algorithm requires computing a coefficient that is based only on local information, as opposed to the globally computed edge betweenness. It is therefore much faster: it runs in $O(|E|^4/|V|^2)$, which is $O(|V|^2)$ for a sparse graph. As a drawback, the method gives worse results on networks where the edge clustering is homogeneously small for all edges, as it happens in some non-social networks [84].

Modularity optimization

A popular and much researched method for community detection is modularity optimization. Modularity is a quality function defined on a partition of a network in communities. It is defined as:

$Q =$ [fraction of edges within communities]-[expected fraction of edges within communities]

A high value of Q indicates that there are more edges in the network inside the partition communities than what would be expected from chance. So, large positive values of Q are expected to indicate good partitions. The expected fraction of edges is evaluated on a *null model* (see section 2.5.1): a graph obtained by randomizing the edges of the network while maintaining the degrees of the nodes. In the null model, the probability of an edge between nodes u and v is:

$$\frac{d_u d_v}{m}$$

where d_u and d_v are the degrees of nodes u and v, and $m = |E|$ is the number of edges in the network. Then, modularity can be defined as:

$$Q = \frac{1}{2m} \sum_{u,v \in V} \left[A_{uv} - \frac{d_u d_v}{2m} \right] \delta(c_u)\delta(c_v)$$

where $(A_{uv})_{u,v \in V}$ is the adjacency matrix of the network, δ is the Kronecker symbol, and c_u is the community to which node u belongs (so $\delta(c_u)\delta(c_v)$ is 1 if nodes u and v belong to the same community, and 0 otherwise).

With this definition, community detection translates to optimizing the benefit function Q over the possible partitions of the network in communities. We can see that this definition does not require as an input the number or sizes of the communities.

Obtaining the optimal value of the modularity is computationally hard: it has been shown that modularity optimization is NP-complete [44]. Therefore, an approximate algorithm is needed in practical situations. The original implementation of Newman and Girvan has a worst-case running time of $O(|E|^2|V|)$, or $O(|V|^3)$ on sparse graphs.

There are different approximate optimization algorithms with a reasonable computation time. We review now a few of them.

- Greedy Algorithms

  A greedy heuristic for modularity optimization was first suggested by Newman in [169]. The algorithm starts with each node being the only element of its own community. In each step, two communities are merged together; the heuristic is to merge those communities that yield a largest increase of the modularity measure Q over the network. If the network has $|V|$ nodes, after $|V|-1$ merging steps, all nodes belong to the same module. Observe that the method naturally gives a hierarchy of modules, that could be represented in a dendrogram. Newman proposed a straightforward implementation of the modularity computation in each step from the adjacency matrix, that gives an overall computation time of $O((|V|+|E|)|V|)$ time, or $O(|V|^2)$ on sparse graphs. The method has shown good performance in some real-world situations.

  Clauset, Newman and Moore [59] proposed a revision of the modularity computation, and used more sophisticated data structures, obtaining a running time of $O(|E| \cdot d \cdot l \cdot \log|V|)$, where d is the depth of the dendrogram built. On sparse networks, that is $O(|V| d \cdot \log|V|)$. Moreover, if the network has indeed a strong hierarchical community structure then the dendrogram is "balanced" and $d \sim O(\log|V|)$, yielding an overall performance of $O(|V|\log^2|V|)$. This makes it theoretically useful for very large networks: up to 10^7 nodes.

  But, if the dendrogram built is not balanced, Clauset et al. algorithm does not perform so well. Wakita and Tsurumi tested the algorithm's performance and suggested the method was practically limited to networks of up to 500 000 nodes. They proposed forcing the algorithm to merge communities in a balanced manner, and showed empirically that their method scales to networks of 5.5 million nodes. It is not clear how realistic this balanced merging is in real-world communities.

  From all modularity optimization techniques, the greedy algorithm is the one able to handle the largest graphs. As a drawback, the approximation to the true modularity maximum is worse than with other techniques [84].

- Simulated annealing. This method is slower, although it can yield results very close to the true maximum. It can be used on graphs of up to 10^4 nodes [84].

- Extremal optimization can be applied to maximize modularity [72]. Starting from a random split, the node with smallest value of a "fitness" function of modularity is moved from one partition to another in each iteration. The algorithm runs in $O(|V|^2 log|V|)$ time for a sparse graph, using efficient data structures.

Observe that modularity cannot be used to compare different graphs. Two graphs with the same modular structure can have different values of modularity: if one is bigger, it will have larger modularity.

A fundamental problem of modularity optimization is the *resolution limit*. Modularity cannot capture communities of a size smaller than a certain scale, that depends on the size of the network and how inter-connected the communities are [84]. If the network is very large, the quality of modularity optimization methods is limited.

Random walks

The intuition behind using random walks to find communities in networks is simple: a random walker would spend more time inside communities than travelling from one to another because, due to the higher density of intra-community links, there are more paths to be followed inside the communities.

As we saw in section 2.5.2, Girvan and Newman proposed a random walker model to give a definition of edge centrality [98]. In their divisive algorithm, edges where the random walker spends less time are identified as likely inter-community edges and removed from the network, in order to disconnect the underlying communities.

Zhou [229] and Zhou and Lipowsky [230] have elaborated on the concept of random walks[54] to detect communities in networks. They define the distance between two nodes as the average number of steps a random walker takes to reach one node from the other. Then, an attractor of a node is its nearest node, using this definition of distance. The intuition behind the community finding method is that every node is with high probability in the same community as its attractor. In particular, in [230] they proposed a hierarchical agglomerative method: starting from each node being in its own community, in every step, the two communities which nodes are on average closer get merged (see [230] for more details on how the distance, called *proximity index*, is defined). This algorithm involves computing the distance between all pairs of nearest-neighbours in the network, and the overall complexity of the method is $O(|V|^3)$.

A different node-to-node distance based on random walks was proposed by Pons and Latapy [183], as simply the probability of a random walker moving from a vertex to another in a fixed number of steps. The worst-case complexity of this method is $O(mn^2)$ time, and space $O(n^2)$, which is also a limiting factor in practical situations. The authors argue, though, that in most real-world cases, the time complexity is in the order of $O(n^2 \log n)$.

Another heuristic approach inspired in random walks is the Markov Clustering Algorithm (MCL) by van Dongen [70]. The method simulates a flow diffusion process in the graph, based on the intuition that a random walk on a dense cluster will probably visit most of the nodes before leaving the cluster[55]. The method involves a costly matrix mul-

[54]Brownian's particle motion in their theoretical physics terminology.
[55]The author has made the code publicly available at http://www.micans.org/mcl/.

tiplication step, so even for sparse graphs, the computational cost is worst-case $O(n^3)$. A problem of the method is that the resulting partition is parameter dependent, and there is no clear way of selecting a solution over another, for different parameter configurations.

Spectral clustering

The popular spectral methods for graph clustering use the eigenvector of the Laplacian matrix of the graph to obtain a partition. The Laplacian matrix of an undirected graph is obtained by subtracting its adjacency matrix from a diagonal matrix of its vertex degrees [15, 54]. Recall that the adjacency matrix of an undirected, simple graph $G = (V, E)$ is $A = (a_{uv})_{u,v \in V}$, where $a_{uv} = 1$ if $(u, v) \in E$ and $a_{uv} = 0$ otherwise. The Laplacian matrix of G is then $Q = D - A$, where $D = (d_{uv})_{u,v \in V}$ is a diagonal matrix in which d_{uu} is the degree $d(u)$ of u in G and $d_{uv} = 0$ for all $u \neq v$. The matrices A and Q are tightly related to several structural properties of the graph [15].

Donath and Hoffman [69] first suggested using spectral methods for graph partitioning. The basic idea is to use the eigenvector of the second smallest eigenvalue of the Laplacian matrix of the graph. In the extreme case that the graph has two equally sized connected components, the sign of the coordinates of the second eigenvector partitions the graph in the two components. In cases where the structure of the graph is less clearly partitioned, it is expected that the second eigenvector still gives a good bisection [15].

Alon [15], Spielman and Teng [204] and Kannan et al. [118] showed that spectral heuristics give indeed good bisections, in terms of coverage and conductance.

Spectral methods take typically polynomial time, so using them directly on very large graphs is not always appropriate [118]. Nevertheless, a number of hybrid techniques have been proposed to speed up the community detection: Drineas et al. [71] use randomization techniques. A drawback of these methods is that they require specifying the number of communities as an input; some adaptations try to find the optimal number of communities automatically [210, 33].

Communities in the Web

In the Web, a community can be defined as a collection of pages that share a common topic [147]. Given the self organised nature of the Web, most of these communities grow spontaneously and are a good indicator of social dynamics of the Web. The obvious approach to look for topical communities would be to locate groups of pages that link to each other. They would be explicitly established communities, where pages, and their authors, are aware of each other. But the social dynamics of the Web give rise to communities even when the participants are not aware of who else is taking part [132].

Flake et al. [82, 83] applied their maximum flow method to find communities on the Web (see above).

Another definition of community in the Web that has been used leverages the concept of *hubs* and *authorities*, as produced by the HITS algorithm [125]. Intuitively, *authorities* are relevant pages for a particular topic, while *hubs* are pages that link to many authorities. The HITS algorithm iteratively updates the authority measure and the hub measure for each of the Web pages: a page is a good *hub* if it links to good *authorities*, and a page is a good *authority* if it is linked by good hubs (see [139]). A previous step to run the algorithm is to obtain, using a search engine, a subgraph of pages relevant to a user query. This subgraph is formed by the (top) documents that contain the query keywords, extended by those pages that link to or are linked from those documents.

Formally: a page i has an *authority score* x_i, and a *hub score*, y_i; starting from a subgraph of the network (that is query dependent in the HITS search algorithm), where e_{ij} is an edge from page i to page j, the HITS algorithm updates iteratively both scores:

$$x_i^{(k)} = \sum_{j:e_{ji}\in E} y_j^{(k-1)}, \quad y_i^{(k)} = \sum_{j:e_{ij}\in E} x_j^{(k)} \qquad \text{for each iteration } k = 1, 2, 3, \ldots$$

Gibson et al. applied this setting to find communities in the Web [96]. Given a query, a subgraph of the WWW relevant to the query is obtained. Then, the HITS algorithm is applied to find hubs and authorities. The top scored are gathered and returned as the "core" community corresponding to the user query.

A method independent of the user query was proposed by Kumar et al. [132]. The authors aimed at finding *emerging* communities, even before the participants are aware of being part of a community . They try to scan the whole Web (a crawl of it, effectively), and find all instances of specific subgraph structures that are the signature of communities (this process is called *trawling* the Web). The main idea is to find communities by *co-citation.* Underlying is the hypothesis that "linkage on the Web represents an implicit endorsement of the document pointed to. While each link is not an entirely reliable value judgment, the sum collection of the links are a very reliable and accurate indicator of the quality of the page" [132].

They develop a mathematical version of this intuition: "Web communities are characterised by dense bipartite directed subgraphs" [132]. A web community would be a dense bipartite subgraph $G = (V = F \sqcup C, E)$ that contains at least one core, where a core is a complete bipartite subgraph with at least i nodes from F and at least j nodes from C.

Our own proposal for community detection in large, directed networks, is inspired by this Kumar et. al work on the Web.

A very interesting study of communities on the Web was performed by Adamic and Adar [3], as it is one of the first analysis of the communities within a Web-based social network. The authors crawled the personal homepages of students of two American universities, Stanford and the MIT. The network was formed by the homepages as nodes and the links between them as edges[56]. The authors went further to analyze the similarity between users, as a function of the co-occurrence of certain items in their homepages. Adamic and Adar suggested this approach to gather social data as an alternative to the labour expensive process of interviewing: the homepages of the students are "proxies" for the individuals; we have seen in section 2.3 how popular this methodology has become.

[56] The network considered was undirected. The authors' intention was that the links between homepages represented that the people linked knew each other. Reciprocity on their dataset was over 50%, suggesting this was often the case. The dataset was small, and the authors were able to remove from their dataset some links that clearly responded to different motives.

A linear time method

Wu and Huberman [224], like Girvan and Newman earlier [98], posed community detection as a current-flow problem. But instead of the computationally expensive procedure of calculating the betweenness of all edges, this is a bipartition method, that can be applied iteratively, based on the physics concept of voltage drop. The resolution of the partition problem scales linearly on the size of the graph. In particular, its complexity is $O(|V| + |E|)$. So it is indeed linear on the size of the graph, and the lowest complexity successful partition method.

The drawback of the method is that it needs a priori knowledge of the number of expected communities, and it assumes that all communities are of similar size [66]. Unlike Girvan-Newman's method, it does not give a hierarchy of communities as a solution.

Graph sampling for community detection

As we have seen, community detection is a computationally hard problem, difficult to tackle for large networks. A suggestive alternative is in sampling. Rather than considering the full network, it would be possible to work only in a representative subset of nodes and edges, and extrapolate the results to the entire graph.

Sampling is a standard statistical approach, but the difficulty is how to find a sample that maintains the community structure of the entire network. Leskovec and Faloutsos influential paper [145] tested on real-world networks the representativeness of samples; they checked, for different sampling approaches, the matching between the sampled graph and the original network. Leskovec and Faloutsos proposed focusing on the degree distribution, the clustering coefficient and the connected component sizes.

Hübler et al. [115] proposed Metropolis sampling algorithms to obtain a subgraph that preserves the degree distribution, clustering coefficient, and the "graphlet distribution": given a graph, a *k-graphlet* is a connected induced subgraph of size k (defined in [158]).

Maiya et al. [152] argue that the former methods of sampling do not keep the community structure of networks. Their model is based on the concept of *expander graphs*, highly connected but relatively sparse graphs. The expansion factor of a subgraph is the ratio of the number of nodes that are neighbours of nodes in the subgraph to the number of nodes in the subgraph. The authors hypothesis is that a subgraph that has good expansion properties is more representative of community structure. They empirically test for theirs and previous sampling methods if the produced subgraphs are representative of community structure. The idea is to run a community detection algorithm, and compare the results on the subgraph and on the entire network. Their results are promising, although it is not clear how sample sizes affect its quality.

Directed networks

It is clear that the direction of the links in endorsement networks is an important information that should not be disregarded. In the Web, hyperlinks are directed. In Flickr, the action of a user u favoring a photo of user a v is qualitatively different from the reverse action, and it is also different from u and v being friends. In food webs, where nodes represent species, linked by predator-prey relationships, the direction of the edges is also very important. As another example, the directionality in gene regulatory networks is also crucial [108]. One of the fundamental goals of the research that we report on this thesis is to take directionality into account.

Algorithms for community detection in directed social networks are relatively new, and the most common practice to deal with directed networks has been to ignore directionality and apply the methods developed for undirected networks. Although, as we have discussed above, many community detection techniques have been proposed, only a few of them can be extended to directed graphs. For example, the successful spectral analysis methods are difficult to adapt as the matrices involved (the adjacency matrix and the Laplacian) are not symmetrical.

A few *undirected techniques* can be adapted to take directionality into account. Leicht and Newman [141] realized that disregarding directions of a directed network and applying modularity optimization can yield to strange results. They reviewed the definition of modularity to generalize it to directed networks. The idea is to look for communities by maximizing modularity over the possible divisions of a network, by using an algorithm based on the eigenvectors of the corresponding modularity matrix.

A particularly interesting approach to the problem of community detection in bipartite networks is the paper of Guimerà et al. [108]. They consider a bipartite graph with *actors* on one side and *teams* on the other, and they propose to optimize a measure of *bipartite modularity*, which adapts modularity to the bipartite case. They suggest the use of the same approach in directed networks, by first projecting the network onto a bipartite graph[57]. As we will see in Chapter 3, our own approach to community detection in directed networks uses a mapping to a bipartite network representation as an instrumental step to discover what we consider the *footprint* of community structure.

Kim et al. [123] also proposed a generalization of modularity for directed networks, by introducing "LinkRank" which is a quantity analogous to PageRank[178], but computed for the edges, rather than for the nodes; LinkRank would measure the importance of an edge as the probability of a random walker following the edge from node i to node j in the stationary state of the random walk. Then, the definition of modularity is modified to be:

$$Q_{LinkRank} = \begin{bmatrix} \text{fraction of time spent walking within} \\ \text{communities by a random walker} \end{bmatrix} - [\text{expected value of this fraction}]$$

Rossvall and Bergstrom also attempt to include the directionality of the links [194] in their previously undirected method [193], based on cluster compression.

Ghosh and Lerman's method [95] is theoretically generalizable for directed networks, although their evaluated implementation is only undirected.

Reichardt et al. [190] proposed a method generalizable to directed and weighted networks. Their method tries to decompose the graph in classes of structurally equivalent nodes. These classes are abstracted to the nodes of a so-called "image graph", that represents the functional roles in the network.

[57] By representing a directed network as a bipartite graphs where the whole set of actors is replicated on the left and on the right side.

Overlapping communities

In many real-world networks, a partition of the network in disjoint subsets of nodes might not be a realistic representation of its community structure. To give an obvious example, a user of Twitter might have a strong interest in more than one topic, say *indie rock music*, *research in social media*, and *racket sports*. It would be natural to identify this particular user as a member of this three "topical communities". As another example, a person might belong to more than one "community of friends", say the community of her college friends, the community of her co-workers and the community of her school friends.

It would be useful then to have methods capable of uncovering the overlapping community structure of a network, where a node can belong, possibly in varying degrees, to more than one group. But most of the definitions of communities that we have seen (see section 2.5.1) yield disjoint groups, and the methods look for partitioning the network in separate groups. Recent research has indeed brought up this research question, but there is still no satisfying solution. Many real networks would be characterised by a significant overlap between different clusters.

Clique Percolation The most common approach to detect overlapping communities is the method of *clique percolation* [179]. The intuition is that communities should have a high internal density, so cliques should be frequently formed. According to this method the definition of the communities is local, based on discovering k-cliques and merging them when they share $k-1$ nodes. The authors have linked this definition to the mechanism of *preferential attachment* [182]. Mathematical proofs of the validity of the method for clique finding were later obtained by Bollobás and O'Riordan [37]

Observe that the idea that edges taking part in cliques are likely within a community, and unlikely inter-community edges, was also exploited by Radicchi et al. in their edge clustering coefficient method [187] (see section 2.5.2).

This approach has been later extended by the same authors to deal with weighted networks [80], and later to directed networks [180] by considering *directed k-cliques*, which are complete sub-graphs of size k in which an ordering can be made such that between any pair of nodes there is a directed link from the higher order node towards the lower one. The authors have implemented their method in a software package, CFinder[58].

This method has two drawbacks. The first is that the algorithm first searches for maximal cliques, which is a problem that grows *exponentially* with the size of the graph, so it is worst-case exponential. The actual average computational cost of the algorithm depends on many factors, and has not been expressed in closed form [84]. The authors, though, report that in real-world networks the method is fast enough to scale to networks of up to 10^5 nodes [179].

The second difficulty with this procedure is that, like the algorithm of Radicchi et al. [187], it assumes a large *cliquishness* of the network, so it may not work so well with less transitive networks, like technological networks"

[58] http://cfinder.org/

Other methods for overlapping community detection Nicosia et al. [175] extended the definition of modularity, considering belonging factors that indicate to which degree each node belongs to each of the communities.

Ahn et al. [9] propose clustering the links instead of the nodes of the network. Their method tries to find overlap *and* hierarchical structure. They define a link similarity and perform agglomerative clustering, obtaining a dendrogram. They use then a quality measure to assess which is the best partition, and cut the tree. The measure is called partition density, and it is basically the average density within the clusters.

Another method working on the links instead of on the nodes of the graph, is the one by Evans et al. [78]. Form the original network, they derive the *line graph*: each edge in the the original graph is represented as a node in the line graph; and there is an edge between two nodes of the line graph if the corresponding edges of the original graph shared and end-point. Evans et al. define dynamic processes, based on random walks, on the line graph, and optimize the associated modularity functions.

The idea of Gregory et al. [102] is to split the overlapping nodes, in as many nodes as clusters they belong to, and then perform any node clustering method. Finally, the partition found is remapped to the original graph, by replacing the split nodes to their original nodes, yielding an overlapping cover.

Gfeller et al. [94], propose a method for identifying those nodes that probably belong to more than one community. They add random noise to the network, and use a conventional method to obtain a clear partition (with no overlap)[59]. Comparing the results of successive runs, they identify the "unstable nodes": those that switch from a cluster to another between different runs. Gfeller et al. define a similarity function between groups of nodes, so they are able to identify stable communities in the network; those are the communities, while the unstable nodes lie in the overlap between these clusters.

The statistical mechanics Q-state Potts model for the interaction between ferromagnetic particles has been proposed for community detection by Reichardt and Bornholdt [189]. The network is mapped to such a model, and it is possible to identify nodes shared between communities, by comparing the partitions produced when looking for global vs. local energy minima.

Finally, Newman and Leicht [172] have suggested using traditional maximum likelihood statistics to compute the probability of a node belonging to a certain community. This method allows for a node belonging to a community to a certain degree, so it is able in principle to deal with overlapping communities. The method is fast, and scales up to networks of 10^6 nodes [84]. The main difficulty with the method is that it requires specifying the number of networks as an input parameter.

2.5.3 Roles, influence and leadership in communities

Once the community structure of a network is revealed, several research questions can be considered. How does the connectivity between communities look like? Are commu-

[59] Any partition method could be applied, in principle. The authors use the Markov Cluster Algorithm of van Dongen [70], described above..

nities isolated or linked together? Are there different roles that nodes play within the communities? How is information and influence spread within a community? etc.

Mesoscopic view After obtaining a clustering of a network in communities, we can consider a derived graph, where each node represents a community and the edges represents the aggregated connectivity among the nodes in different communities. This is known as the *mesoscopic view* of the network. A first conclusion of the study of the inter-community connectivity of real-world networks is that the size distribution of communities is skewed[85]. Clauset et al. [59], for instance, studied the Amazon co-purchasing network[60]; they showed that the sizes of the communities found by modularity optimization follow a power-law. Similar results have been obtained for other networks[21, 169, 179]. Pollner et al. [182] have suggested that the same mechanism of *preferential attachment* that produces long-tail degree distributions is also behind the distribution of community sizes.

Influence In social groups, individuals influence each other's behaviour: for instance, new social practices spread through society and word-of-mouth is considered to play an important role in shaping consumers' attitudes [48]; and viral marketing is a successful new form of advertising [143]. Understanding why and how influence operates on society are central questions of social sciences. Assuming the structural hypothesis of social network analysis, to explain the phenomenon of influence in networks we should look at the network of interactions between individuals.

There are a variety of reasons that have been suggested in order to explain influence in social networks[61].

A first explanation is based on the idea that people assume "the behaviour of others conveys information" [74]; it has been argued that, when taken decisions, we often assume that people who made a particular choice before us did so because they had information that we might not have. For instance, if a Web site is used by many people, we assume this as a signal of its quality.

A different explanation is based on the idea that people perceive that aligning their behaviour with that of others produces a social benefit [74]. For example, there are a variety of competing Social Network Sites; as they are used for *connecting* with others, and *sharing* media, the more users the site already has, the more valuable is for the user to join it. It is clear the analogy with the process of *preferential attachment*: the rich-gets-richer effect.

Spreading of rumours, information and diseases is one of most studied effects of influence in social networks. Influence through a network can indeed have a "cascading effect": Leskovec et al. studied how email recommendations for a graphic novel spread through the network, and described this effect as "social contagion" [143].

[60] In this network, the nodes represent Amazon customers. A link between two customers represents that they bought the same product.

[61] A thorough discussion of these topics can be found in the books by Easley and Kleinberg [74] and Wasserman and Faust [215].

An interesting problem is to identify the most influential individuals, those which messages are spread further and wider in the social network. In social network analysis, the power of a node for spreading has been most often measured by its *betweenness centrality*: how many shortest paths across the network contain the node; those nodes with a higher centrality are in a position to control communication on the network. Another common notion of importance in the network is that the most connected nodes (*hubs*) have the largest spreading power [14]. Both notions are intuitive, but a very recent study suggests that the picture is not so simple. Kitsak et al. [124] suggest that is some situations, the highly connected or highly central nodes might not possess a large spreading power: if a *hub* is located at the periphery of the network, it will have less impact that a less connected node strategically situated at the core of the network. These authors use *k-shell decomposition* [53] to identify the global *nucleus* of the network[62]: a small group of nodes that form an extremely well connected distributed subgraph. The decomposition is performed by recursively pruning the least connected vertices, so in each step the best connected nodes are kept, and those removed in step k are assigned to the k-shell [63]. Simulating rumour and disease spreading models in the network, they validate the intuition that the *k-shell level* is a good indicator of the spreading influence, which is highest for those few nodes in the nucleus. As we will see in Chapter 3, our own intuition is to find *local nuclei*, and consider that those nodes in a nucleus have a higher spreading power.

Mathioudakis and Koudas [154] define, in the context of the *blogosphere*, “starters”, which are bloggers that generate content and receive links from others; and “followers”, which are users who mostly comment on and link to content generated by other users, while not receiving comments or links themselves. We also, as we shall see, propose the definition of “leaders” and “followers” in networks, although our definition is strictly based on linking patterns, and we do not limit the volume of attention a follower can receive (see Chapter 3).

Roles The position of nodes within communities can be considered indicative of the *role* they play in the network. Nodes that are central in the community, having a large number of links to other nodes in the group, would have important functions of control within the group [84]; nodes that lie in the borders of communities would play the role of inter-community communication.

Position and roles in social networks have been studied for a long time by social scientists (see [215]). The methodological assumption underlying network analysis of individual roles is that the role of an individual in society can be described, and measured, by the ties they have to other individuals. The goal is to find similar patterns of interaction of individuals and use them to describe the roles of social actors.

Starting from the uncovered community structure of a network, Guimerà et al. [105, 104] proposed a classification of the nodes within the communities according to their roles. Guimerà et al. propose a classification of “universal roles” for nodes, based on

[62] Carmi et al. [53] use the term *nucleus*, that we prefer over *core*, used by Kitsak et al. [124].
[63] The theoretical foundations of this technique lay within percolation theory; see section 2.4.2.

clustering them around the values of two measures: (1) the relative module degree, z, which is a measure of how well connected a node is to the rest of the community, relative to the connectivity of these other nodes; and (2) the participation coeficient, D, which quantifies to what extent a node *connects* different communities. Later, the same authors proposed a classification of complex networks according to role-to-role connectivity [107].

These ideas are also coherent with social sciences theories: the *weak ties* theory of Granovetter [101] and Burt's *structural holes* [50], that we have already mentioned in section [101].

2.5.4 Evaluation

The concept of community is intuitive, but there is no unique or obvious definition of what a community is. On a qualitative level, it seems clear that the members of a community will, with a high probability, share common features or play similar roles within the graph[83]. On a quantitative level, in order to obtain the communities of a network, the first step is to quantitatively specify a definition of community (see section 2.5.1). From the definition, an algorithm should be able to obtain all communities in the network. But, even for small networks, this task may be computationally unfeasible[187]. The solutions algorithms offer is consequently limited, an approximation. And the quantitative comparison of the results of different algorithms is not a clearly defined task. A common approach is to "check whether the results appear sensible"[187]. This is, obviously, no definitive solution.

Qualitative evaluation We mentioned in section 2.5.2 the seminal research of Adamic and Adar of communities in the network of personal homepages of people associated with two American universities [3]. The authors not only studied the link structure, but also wanted to qualitatively analyze the individuals whose homepages formed the network. Their intuition, based on the observation, confirmed by social sciences research, that "friends tend to be similar"[3][64]. They considered four kinds of information: text (in the form of noun phrases extracted from the homepages), links from the homepage (*out-links*), links to the homepage (*in-links*), and links from the homepage to mail-lists. They proposed computing the similarity between users A and B as a function of the information items their two homepages shared. In particular:

$$similarity(A, B) = \sum_{shared\,items} \frac{1}{log(frequency(shared\,item))}$$

So items are weighted inversely respect to their frequency in the dataset: infrequent items are assumed to be a better indicator of user similarity. The authors ranked users according to their similarity, and evaluated, for a user, if highly ranked users where "friends": they were explicitly linked from the user's homepage. They compared the

[64] In social science research, this has been explained through feedback mechanisms. People who interact tend to share information and knowledge, while the relative sharing of knowledge between two people makes interaction more likely[52].

predictive power of the four information sources and concluded that *inlinks* are the most predictive. They also give some empirical evidence that certain items are more correlated with friendship than others: too general items are bad predictors for social connections.

The same authors extended their analysis to the study of one of the first Social Network Sites, "Club Nexus", a community of Stanford students that is an antecedent of the current Orkut.

Benchmark graphs A possible approach for evaluating community detection methods is to use benchmark graphs. A benchmark graph is a synthetic graph generated by some defined random process, that is supposed to replicate the important statistical features of real-world networks. A benchmark graph that has been re-used in the literature was proposed by Girvan and Newman in [98] to test their community finding method. A network generated according to their model will have 128 nodes, divided in four groups of 32 nodes. As in a random graph, an edge between two nodes exists with a certain probability, independently at random. But the probability of an edge joining two nodes of the same group is higher to the probability of joining two edges in different groups. The four groups of nodes are assumed to have "well defined community structure".

The Girvan-Newman synthetic graph is rather unrealistic. Like in a random graph, most nodes have the same degree. Furthermore, all communities are equally sized, there is no hierarchic structure, and the inter-community connectivity is homogeneous. Lancichinetti et al. argued that the Girvan-Newman benchmark graph is not adequate to test the reliability of algorithms on real-world networks[137], and proposed more elaborate benchmark graphs.

In [137], Lanchichinetti et al. propose a benchmark graph where the node degree distribution and the community size both follow a power-law, where the exponents are parameters of the model. In [135], they extend their model to account for overlapping communities, directed and weighted networks. The same authors have proposed recently a measure of similarity that can be used to compare two divisions of a network into overlapping communities[136].

A similar approach to generating benchmark graphs is to generate random graphs that match a real-world network in a number of identified statistical features, that are feed to the model by means of parameters: "measurement-calibrated graphs"[196].

2.6 Conclusions

We have covered a wide area of research in this section. In particular, a great variety of community detection approaches has been discussed in section 2.5. We give now a few concluding remarks, that will inform our own investigation, to be discussed in Chapter 3.

- As it is for any particular problem of network analysis, research on community detection is interdisciplinary. Valuable observations, results, algorithms and experiments have been proposed from the social sciences, physical sciences and computer science, among others.

- There have been a great number of methods developed. The scalability of the methods has improved, but the problem is still hard. The Table 2.1 summarises the computational complexity of some of the most important community detection methods we have reviewed. As the size of the datasets that interest scientists grows, many of the best methods of community finding are simply not applicable. The most popular family of methods, modularity optimization, has been optimized thoroughly; despite that, it is not usable for very large graphs (in the order of millions of nodes): at present the fastest method for modularity optimization is $O(n \log^2 n)$, but it does not guarantee the best partition, or indeed a very good one, if the network community structure does not have certain properties (see [59, 213, 66]).

- Complexity is not the only factor to take into account when evaluating the usefulness of a community finding method. As discussed before, each method has particular limitations. When deciding if a particular method gives "good" partitions, it is important to consider the particular kind of networks we are interested in.

- There are a few natural aspects of community structure that the current state of the art does not address satisfactorily; in particular, hierarchical and overlapping communities, and directed networks. Many of the methods work on undirected networks only. The usual approach to directed networks has been to neglect the directions of the edges, and work on the underlying undirected graph. Some undirected methods have been adapted to consider directionality, but the quality of the solutions has not been tested as thoroughly as for undirected networks. Fortunato and Castellano describe the ideal method as one that [84]: "delivers meaningful partitions, and handles overlapping communities and hierarchy, possibly in a short time". But "No such methods exist yet". We would add that these methods should also handle directionality satisfactorily.

- The lack of a common evaluation framework makes comparisons between algorithms difficult, although there have been some recent developments in that sense. Moreover, the task of community detection, despite being intuitively clear, is not defined in clear terms that are shared by the wealth of literature published on the topic.

Table 2.1: Summary of the worst-case computational complexity of community detection methods for a network $G = (V, E)$ with $|V| = n$ nodes and $|E| = m$ edges.

Algorithm	**Authors and references**	**Complexity**	**Observations**
PARTITIONING			
Kernighan-Lin	Kernighan and Lin [121]	$O(n^2)$	See also [166, 43, 231]. Both are heuristic, very dependent on the initial configuration, and tend to fall in local maxima.
Kernighan-Lin	Fiduccia and Mattheyses [81]	$O(m)$	
DIVISIVE			
Edge betweenness	Girvan and Newman [98]	$O(m^2 n)$	
Information centrality	Fortunato et al. [86]	$O(m^3 n)$	
Edge clustering coefficient	Radicchi et al. [187]	$O(m^4/n^2)$	
MODULARITY			
Newman-Girvan	Newman and Girvan [171]	$O(m^2 n)$	
Greedy heuristic	Newman [169]	$O((m+n)\,n)$	
Clauset, Newman and Moore	Clauset, Newman and Moore [59]	$O(m\,d\,\log n)$	$d =$ depth of the dendrogram built For sparse graphs with balanced hierarchical community structure: $O(n\,\log^2 n)$. See also [213]
Extremal optimization	Duch and Arenas [72]	$O(n^2\,\log n)$	
RANDOM WALK			
Zhou and Lipowski	Zhou and Lipowski [230]	$O(n^3)$	
Pons and Latapy	Pons and Latapy [183]	$O(mn^2)$	Space complexity $O(n^2)$
OTHER			
Wu-Huberman	Wu and Huberman [224]	$O(m+n)$	Input number of communities. Similar size communities.
Clique Percolation Method	Palla et al. [179]	$O(\exp(n))$	Overlapping communities. Adaptations for directed and weighted networks [180, 80].

3 Experiment: Coalescing Cores into Communities

3.1 Definitions

We denote an endorsement network by $G = (V, E)$, where V is a set of nodes and E is a set of *directed* edges. A directed edge $(u, v) \in E$ indicates an action of endorsement from node u to node v.

The semantic definition of endorsement depends on the network under consideration, for instance, in a photo-sharing site an edge (u, v) may signify that user u *likes* at least k photos of user v, while in a micro-blog site an edge (u, v) may signify that user u *follows* user v.

We define $N_{\text{in}}(u) = \{v \mid (v, u) \in E\}$ to be the set of *incoming* neighbors of u and $d_{\text{in}}(u) = |N_{\text{in}}(u)|$ to be the *in-degree* of u. Similarly we define $N_{\text{out}}(u) = \{v \mid (u, v) \in E\}$ to be the set of *outgoing* neighbors of u, and $d_{\text{out}}(u) = |N_{\text{out}}(u)|$ to be the *out-degree* of u.

Definition 1. Density. Given two sets of nodes $A, B \subseteq V$, $A \cap B = \emptyset$ we define the *density* $\delta(A, B)$ of the set A towards the set B to be the fraction of the number of all edges from nodes in A to nodes in B, over the number of all possible such edges. That is,

$$\delta(A, B) = \frac{|\{(u, v) \in E \mid u \in A \text{ and } v \in B\}|}{|A| \cdot |B|}.$$

Definition 2. Internal density. We define the *internal density* $\delta_{\text{int}}(A)$ of the set A, as the fraction of the number of all edges between nodes in A over the number of all possible edges in A. That is,

$$\delta_{\text{int}}(A) = \frac{|\{(u, v) \in E \mid u \in A \text{ and } v \in A\}|}{|A|(|A| - 1)}.$$

Definition 3. External density. Similarly, we define the *external density* of the set $A \subseteq V$, as $\delta_{\text{ext}}(A) = \delta(A, V \setminus A)$, i.e., the fraction of the number of all edges from A to $V \setminus A$ over the number of all such possible edges.

Central to our study is the concept of *core*.

Definition 4. Core. Let $G = (V, E)$ be an endorsement network. A *core* $C = (L, F)$ of the network G consists of two disjoint subsets of V, i.e., $L, F \subseteq V$ with $L \cap F = \emptyset$, so that for each $u \in F$ and $v \in L$ it is $(u, v) \in E$. The set L represents the *leaders* of the core, and set F represents the *followers* of the core. The set of leaders L is also called the *nucleus* of the core.

According to the above definition, for each follower in the core there are edges to all the leaders in the core.

Definition 5. Size and support of a core. Given a core $C = (L, F)$, we define the *size* of the core $s(C)$ to be the size of the leader set L, i.e., $s(C) = |L|$, and the *support* of the core $\sigma(C)$ to be the size of the follower set F, i.e., $\sigma(C) = |F|$.

Given a core $C = (L, F)$, we define the *leader-leader density* of the core $\delta_{\mathrm{LL}}(C)$ to be the internal density of the leader set L, and the *follower-follower density* $\delta_{\mathrm{FF}}(C)$ to be the internal density of the follower set F. That is, $\delta_{\mathrm{LL}}(C) = \delta_{\mathrm{int}}(L)$, and $\delta_{\mathrm{FF}}(C) = \delta_{\mathrm{int}}(F)$. We also denote $\delta_{\mathrm{LF}}(C) = \delta(L, F)$ and $\delta_{\mathrm{FL}}(C) = \delta(F, L)$. Later we will use the same notation for communities. Note that for cores $\delta_{\mathrm{FL}}(C) = 1$ by definition.

3.2 Dataset Analysis

3.2.1 Datasets

We analyze five datasets, three endorsement networks and two social (i.e., not endorsement) networks.

Flickr endorsement network (Flickr-E). Flickr is a popular photo-sharing social network. Flickr users can mark photos of other users as *favorites* or they can make *comments* to photos. Marking a photo as favorite is clearly an action of endorsing authority, and in practice so it is making a comment, as most of the comments are praise of the skills of the photographer ("Nice shot!", etc.). We sample a subset of the entire Flickr social network by applying the snowball sampling strategy, starting from a single seed user and following the contact links between users in an iterative manner. Thus, we generate an endorsement network by considering a directed edge between two users u and v if user u has marked at least one photo of user v as *favorite* or if s/he has made at least one *comment* in a photo of v.

Jaiku micro-blog network (Jaiku). Our second endorsement network is Jaiku,[1] which is a social networking and *micro-blogging* service, comparable to the better known Twitter[2]. Jaiku allows users to post short status messages and other content. Users

[1] www.jaiku.com

[2] We decided to use Jaiku instead of Twitter because of the difficulty of gathering a comprehensive Twitter dataset. Given the huge size of Twitter, and the restrictions to crawling imposed by the system at the time of performing these experiments, it was difficult to obtain a group of nodes for which we know most of the connecting edges. Twitter, like the Web and other networks, has a giant

can establish contact with each other by subscribing to their feed (*following*). We can therefore build an endorsement network where nodes are users and there is a directed edge from user u to user v whenever user u is following user v. We have conducted a crawl via Jaiku's public API, starting from a set of 632 users, obtained from the results of a popular search service. The dataset can be made available for academic research upon request. This process yielded a network with 31534 nodes. It is important to note that a user can decide to keep his feed and contacts information private. In our network, these users are dangling nodes, since no information of the edges coming out of them can be extracted. About 18.5% of the users we inspected were private users.

Epinions trust network (EPINIONS). This is a who-trust-whom on-line social network of a the general consumer review site Epinions.com. Members of the site can decide whether to "trust" each other. All the trust relationships interact and form the a so called "Web of Trust" which is then combined with review ratings to determine which reviews are shown to the user. It is clearly an example of endorsement network.

Flickr social network (FLICKR-S). Flickr also allows the user to create a social network by declaring other users as *contacts*, *friends*, or *family*. We use the same Flickr user sample as in FLICKR-E to extract a social network among users. Now, a directed edge from u to v indicates that user u has marked user v to be their *"friend"* or *"family"* (and not simply *"contact"*). Notice, that since in Flickr users do not have to reciprocate the friendship or family links, the FLICKR-S network is directed. Since the links indicate social relationship and do not endorse authority among users, we acknowledge this network to be a pure social network and not an endorsement network.

Yahoo! 360 social network network (Y!360). The second social network we use is Yahoo! 360,[3] a personal communication portal that included features such as creating personal web sites, photo sharing, blogging, reviewing products, and more. The dataset we use is available for academic research through the Yahoo! webscope program.[4] Our Y!360 dataset is an undirected network that indicates friendship relationship among users, so it is not an endorsement network. This is the only undirected dataset that we consider, but obviously we can consider links in the two directions, yielding a directed network.

3.2.2 Dataset statistics

The basic characteristics and statistics of our datasets are reported in Table 3.1. Notice that JAIKU and EPINIONS are significantly smaller. On the other hand, the Y!360

connected component. [134], that we would have need to crawl; being in the order of millions of profiles, it was not feasible for us. We leave working on Twitter data for future work. We have discussed the challenges to data availability, in section 2.4.4.

[3] Yahoo! 360 officially closed on July 13, 2009

[4] webscope.sandbox.yahoo.com

Table 3.1: Network Statistics. n: number of nodes; m: number of edges; $\bar{d}$: average degree; $\max d_{\text{in}}$: maximum in-degree; $\max d_{\text{out}}$: maximum out-degree; R: reciprocity; α_{in}: exponent of the power-law of the in-degree distribution; α_{out}: exponent of the power-law of the out-degree distribution; max CC: size of the largest (strongly) connected component; |CC|: number of the (strongly) connected components; c: clustering coefficient.

Network	Flickr-E	Flickr-S	Jaiku	Epinions	Y!360
n	826 829	687 091	31 534	75 879	1 921 351
m	65 851 110	10 122 046	231 006	508 837	7 230 996
$\bar{d}$	79.6	14.7	7.3	6.7	3.8
$\max d_{\text{in}}$	22 214	7 610	2 324	3 035	260
$\max d_{\text{out}}$	15 090	2 867	48	1 801	260
R	0.21	0.48	0.44	0.25	1.00
α_{in}	1.6	2.1	1.7	1.6	2.5
α_{out}	1.7	1.8	1.1	1.7	2.5
max CC	486 210 (58.80%)	479 127 (69.73%)	21 937 (69.57%)	32 223 (42.46%)	1 463 264 (76.16%)
|CC|	341 604	334 933	17	42 185	150 773
c	0.08	0.04	0.06	0.07	0.03

network is the sparsest. We note that even though the networks Flickr-E and Flickr-S are defined over the same base of users, they have different set of nodes; the reason is that singleton nodes have been removed.

As expected, in all networks the in-degree and out-degree distributions follow power laws. The exponents of the distributions, α_{in} and α_{out} are shown in Table 3.1. $\bar{d}$ denotes the average in-degree and out-degree, which have to be equal.

For a directed network we define *reciprocity* to be the fraction of edges that are reciprocal. Obviously for Y!360 reciprocity is 1. Notice that for the two Flickr datasets, which are networks over the same set of users, the social network Flickr-S has much higher reciprocity than the endorsement network Flickr-E.

Finally, we note that the clustering coefficient values have been computed considering all edges as undirected, and approximating by sampling for the larger networks Flickr-E, Flickr-S and Y!360.

3.3 Mining Cores

The first technical problem we face is how to mine cores in the endorsement network G. We first observe that single nodes induce uninteresting trivial cores, such as $(\{u\}, N_{\text{out}}(u))$ and $(N_{\text{in}}(u), \{u\})$. In order to mine more interesting cores, we resort to size constraints, namely we set a lower bound constraint on the size and the support of cores.

Problem 6. (Mining cores) Given an endorsement network G, a threshold value s_0 on core size, and a threshold value σ_0 on core support, we seek to find all cores C in G that have size $s(C) \geq s_0$ and support $\sigma(C) \geq \sigma_0$.

It is immediate that Problem 6 is an instance of the frequent-itemset mining problem [8]. Recall that in the frequent-itemset mining problem, we are given a set of transactions, each transaction being a set of items, and the task is to find all itemsets than co-occur in at least k transactions. The mapping of the core-mining problem to the frequent-itemset mining problem is quite straightforward: each node u in the network G represents a transaction $t(u)$, and $t(u)$ contains all nodes in the set $N_{\mathrm{out}}(u)$. A core $C = (L, F)$ then corresponds to a frequent itemset X found by the algorithm. The nucleus of leaders L corresponds to the itemset X and the set of followers F corresponds to the transactions that support the itemset X. Thus, mining all the frequent itemsets with size greater than s_0 and support greater than σ_0 gives all cores required for Problem 6.

As usually happens with frequent itemsets, the result set is likely going to contain many cores (obviously depending on the selectivity of s_0 and σ_0). Among the various strategies to deal with the patterns explosion problem, an interesting one is to consider only *maximal frequent itemsets* [30]. A maximal frequent itemset is simply an itemset which is frequent and has no frequent superset. In our context this means that given a minimum number of followers σ_0 we are not interested in a core where the nucleus of leaders is X, if the nucleus $X \cup \{v\}$ has still enough followers. Thus the benefit of extracting only the maximal nuclei is twofold: (i) fewer and more interesting cores, and (ii) more efficient computation.

3.3.1 Preprocessing

In order to reduce the size of the mining problem, we perform some preprocessing steps: first, we perform a data-reduction step that removes all nodes u from being considered candidates for leaders or followers, whenever they have $N_{\mathrm{in}}(u) < \sigma_0$ or $N_{\mathrm{out}}(u) < s_0$, respectively. Since the removal of nodes decreases the in-degree and out-degree of other nodes, the data-reduction preprocessing can be repeated iteratively until a fixed-point is reached, without losing any valid solution [133, 38].

We also remove from candidate leaders the nodes with in-degree larger than a given threshold $d_{\max}$. For our experiments we set $d_{\max}$ to be 1% of the total number of nodes in the network. This pruning accelerates computations dramatically, while removing those outliers and letting us focus on nodes that are not *too popular*. The intuition behind this pruning step is to not consider in the analysis top professional photographers in a context like Flickr, or big media in a context like Twitter, and instead focus on the "standard" users. Moreover the top in-degree nodes can in any case be easily identified and their community of influence extracted. We aim instead at those that are hidden in the vast amount of data. Note that nodes with in-degree above the threshold could still be potential interesting *followers*.

We have decided to use the software for frequent itemset mining by Uno, Asai, Uchida, and Arimura: LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets[5]. This program has nice qualities: it is almost memory-constant throughout the calculation, and outputs not only the frequent items identifiers but also the transaction identifiers,

[5] `http://research.nii.ac.jp/~uno/code/LCM.htm`

that in our case are the fan nodes. The algorithm takes linear time to preprocess, and the computation time is dependent on the number of itemsets found, and the minimum support parameter. Rougly speaking, if the ouput produced has size equal to the input database, producing each itemset takes constant time. Overall, the method is considered to be very efficient, and in our experiments has always taken minutes for reasonable selections of the support parameter.

To recap, our approach consists of: (1) representing the *directed* network as a *bipartite* graph; (2) prune the network so we reduce the size of the problem; (3) mine the network for bipartite *frequent structures* (cores).

3.3.2 Empirical observations

The results of various cores extractions are reported in Table 3.2 and Table 3.3. It is worth mentioning that we can not use the same settings of the parameters s_0 and σ_0 in all the networks, as they have different sizes and different densities: what is a reasonable settings for one network could result in too few cores in another network. The table reports statistics for all the 5 networks: three endorsement (social) networks, Flickr-E , Jaiku and Epinions); and two purely friendship-based social networks, Y!360 and Flickr-S.

Various observations can be drawn from Table 3.2.

1. Many large cores can be found in both endorsement and social networks. As it is obvious, this number depends on the settings of the s_0 and σ_0: as the thresholds are higher, less cores are found. Core leader-leader density $\delta_{\mathrm{LL}}(C)$ is always much larger than the follower-follower density $\delta_{\mathrm{FF}}(C)$: as shown in Figure 3.1 the probability that a node links to another node in the same core is usually some orders of magnitude larger than the probability that it points to a node outside the core.

2. The number of nodes which are follower in at least one core (i.e., $|\mathbb{F}|$), is usually one order of magnitude larger than the number of nodes which are leader in at least one core (i.e., $|\mathbb{L}|$).

3. A large number of the nodes which are leaders in one core, happen to be follower in another[6] core (usually in between 50% and 95%). This fact might be interpreted positively, saying that influential users are also followers of other users, thus viral propagation of information is indeed likely.

4. The average density of the nuclei does not depend on the number of cores found. It also does not depend on the minimum support threshold σ_0, while it seems to depend on the minimum size of the nucleus s_0. In particular a higher value of s_0 induces an higher density. This means that as nucleus are usually very dense, missing only one or two of the possible links, this few missing links degrade more the density value in smaller nuclei, than in larger nuclei.

[6] Notice that a node can never be both follower and leader in the same core. In fact, to be follower a node must follow *all* the leaders in the core (i.e., the core is a biclique), but a node can never be follower of itself by definition (i.e., our endorsement directed networks never contain self-loops).

5. In both endorsement and social networks, the average density of links among the followers (i.e., δ_{FF}) is always much lower than the nucleus density (i.e., δ_{LL}). This clearly shows the presence of a strong *directionality* of the links: mainly from the followers to the leaders. Recall that $\delta_{\mathrm{FL}}(C) = 1$ by definition, or in other terms, in a core all followers point to all leaders.

6. The most important empirical observation we obtained: leaders in a core endorse each other forming a very dense nucleus. In most of the settings, in endorsement networks the internal density of the nucleus of a core (i.e., δ_{LL}) is above 87%. This density is evidently lower in friendship-based social networks. Only in Flickr-S for a minimum size $s_0 = 6$, the internal density of the nuclei becomes very high, also due to the low number of cores. In any case, not as high as for Flickr-E. A density distribution in the two different datasets is shown in Figure 3.2(a) and (b).

Table 3.2: For various values of and (columns 1-2): numbers of cores found (column 3); total number of nodes which are follower (respectively, leader) in at least one core, i.e., , and (columns 4-5); number of nodes that are leader in one core and follower in another one (column 6); average leader-leader and follower-follower density (columns 7-8).

Flickr-E

s_0	σ_0	# cores	$\lvert\mathbb{F}\rvert$	$\lvert\mathbb{L}\rvert$	$\lvert\mathbb{F}\cap\mathbb{L}\rvert$	avg δ_{FF}	avg δ_{LL}
4	120	65 868	10 806	653	551	0.41	0.80
4	150	5 777	4 974	198	174	0.37	0.82
5	90	928 484	9 631	876	731	0.54	0.87
5	100	264 548	7 303	585	485	0.51	0.87
6	90	630 476	4 614	442	362	0.59	0.92
6	100	145 298	3 106	241	222	0.56	0.92

Epinions

s_0	σ_0	# cores	$\lvert\mathbb{F}\rvert$	$\lvert\mathbb{L}\rvert$	$\lvert\mathbb{F}\cap\mathbb{L}\rvert$	avg δ_{FF}	avg δ_{LL}
4	50	5 813	722	96	94	0.47	0.89
4	60	1 100	544	56	54	0.43	0.92
5	50	1 364	368	42	41	0.49	0.92
5	60	64	224	17	17	0.43	0.94
6	50	49	163	17	17	0.48	0.92

Table 3.3: For various values of and (columns 1-2): numbers of cores found (column 3); total number of nodes which are follower (respectively, leader) in at least one core, i.e., , and (columns 4-5); number of nodes that are leader in one core and follower in another one (column 6); average leader-leader and follower-follower density (columns 7-8).

Jaiku

s_0	σ_0	# cores	$\lvert\mathbb{F}\rvert$	$\lvert\mathbb{L}\rvert$	$\lvert\mathbb{F}\cap\mathbb{L}\rvert$	avg δ_{FF}	avg δ_{LL}
5	50	242	258	42	12	0.48	0.89
5	30	11 248	356	97	53	0.59	0.87
4	50	286	449	48	14	0.44	0.86
4	30	13 748	1 252	193	121	0.59	0.86
3	50	385	1 625	107	48	0.36	0.77
3	30	15 292	2 829	346	243	0.57	0.84

Flickr-S

s_0	σ_0	# cores	$\lvert\mathbb{F}\rvert$	$\lvert\mathbb{L}\rvert$	$\lvert\mathbb{F}\cap\mathbb{L}\rvert$	avg δ_{FF}	avg δ_{LL}
4	150	2 010	2 087	110	67	0.40	0.79
4	120	29 492	4 431	351	243	0.43	0.60
4	90	836 479	7 443	930	668	0.46	0.48
4	120	29 492	4 431	351	243	0.43	0.60
5	90	247 021	3 474	426	288	0.52	0.69
5	100	69 545	2 506	269	170	0.50	0.76
6	80	456 110	2 118	311	192	0.57	0.80
6	120	1 583	512	35	33	0.48	0.90

Y!360

s_0	σ_0	# cores	$\lvert\mathbb{F}\rvert$	$\lvert\mathbb{L}\rvert$	$\lvert\mathbb{F}\cap\mathbb{L}\rvert$	avg δ_{FF}	avg δ_{LL}
4	50	8	109	8	4	0.29	0.62
4	40	66	262	25	11	0.33	0.70
5	40	1	43	5	0	0.31	0.50

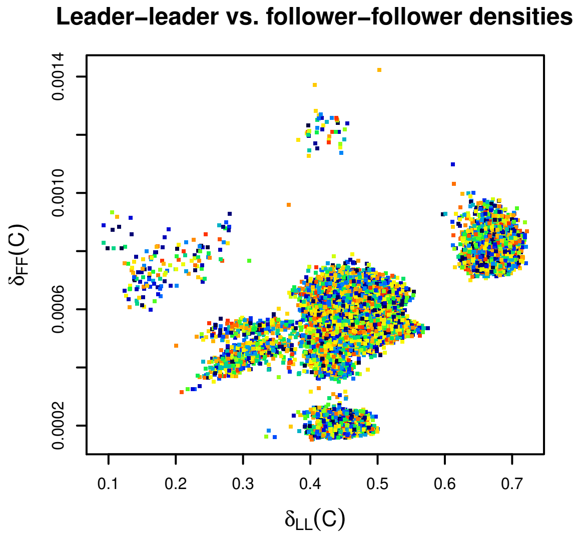

Figure 3.1: Scatter plot of core leader-leader versus follower-follower density of FLICKR-E.

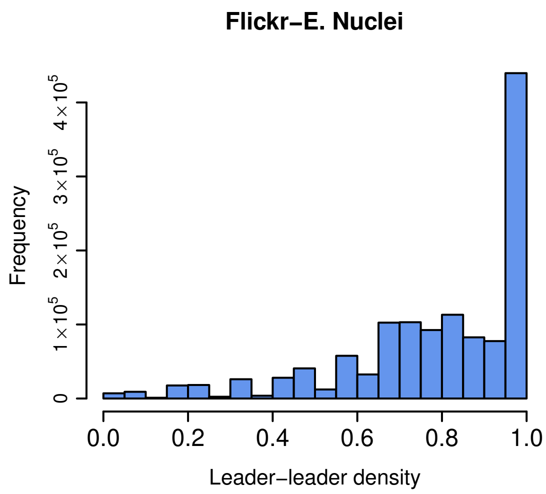

Figure 3.2: Histogram of the nucleus internal density (i.e., δ_{LL}) of the cores found with $s_o = 4$, $\sigma_0 = 90$ in FLICKR-E.

3.3.3 Statistical significance

We present in this section the statistical significance tests that we have performed to check that the structures that we hypothesize are the footprint of communities do not appear by chance.

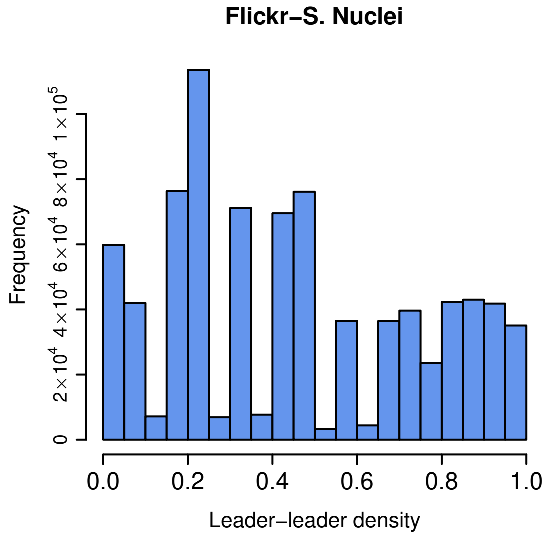

Figure 3.3: Histogram of the nucleus internal density (i.e., δ_{LL}) of the cores found with $s_o = 4$, $\sigma_0 = 90$ in Flickr-S.

First, we present an intuitive test that shows that the internal density in the nuclei is not produced only by the high in-degree of the participant nodes.

Then we use swap randomization, which is a sound randomization scheme, to validate the significance of the findings about cores *and* their densities.

Is internal density produced by high in-degree?

We have come across the unexpected discovery of the sets of leaders (*nuclei*) being very dense internally. In fact, many of them are completely connected. The nodes that take part in these sets, the *leaders*, have some peculiarities: in particular, they have high in-degree. We could imagine that the high density of the nuclei is just a consequence of the special properties of the leader nodes.

A very simple way of deciding whether the internal linkage between leader nodes is just a consequence of the qualities of the individual nodes involved is to take random sets of centers and see if the internal density is preserved. We report on the results of this experiment only for the Flickr-Edataset. The swap randomization scheme used below gives a stronger proof of the validity of the findings.

In the output of the maximal frequent itemset mining with minimum support 100 and minimum size 5, we have found 264 548 different sets, formed from combinations of the 585 nodes that take the role of leaders in some of them. The leaders are nodes with a high indegree: $median = 566$ and $mean = 583.4$. We want to test that this is not the only reason why they form dense nuclei. We produce an equally large sample of random combinations of the 585 nodes, and compute their densities. For simplicity, we only produce sets of 5 nodes [7].

That is: we produce a sample of 264 548 random sets, which elements are taken from the 585 centers, and compute the average densities of the sets. We have repeated the sample 10 times, to assess significance, and have obtained consistent results across them. While the nuclei found where very dense internally, with mean internal density:

$$avg\,\delta_{int} = 0.87$$

The random 5-nodes leader sets are not internally dense. Averaging through 10 runs of choosing random 5-nodes center sets:

$$avg\,\delta_{int} = 0.068854$$

Swap randomization

In our experimental evaluation, we validate the statistical significance of our findings using the method of *swap randomization* [97]. We give here a brief overview of the approach.

Let G be a network, for which assume that we have executed a data-mining algorithm $\mathcal{A}$, and we have obtain aggregate result $X_0 = \mathcal{A}(G)$. One way to test the significance of result $\mathcal{A}(G)$, is by generating k *random* networks $G_1, \ldots, G_k$, such that each network G_i, $i = 1, \ldots, k$, not only has same nodes as G but also each node u has the same degree in G and G_i. For the method to work, it has to ensure that G_i is a network drawn uniformly at random among all networks that have the same *degree sequence* with G. Such a uniform random sample G_i can be obtained by performing a random walk on the space of all networks with the same degree sequence. By performing sufficiently many steps of the random walk and appropriately rejecting steps according to the Metropolis-Hastings algorithm [110, 156] it is guaranteed that the stationary distribution of the random walk is the uniform one.

The algorithm $\mathcal{A}$ is executed on each sampled network G_i, yielding results $X_t = \mathcal{A}(G_i)$ for $i = 1, \ldots, k$, and the significance of the result $\mathcal{A}(G)$ of the algorithm $\mathcal{A}$ on the network G is tested by comparing it against the set $\mathbf{X} = \{X_1, \ldots, X_k\}$ of the results of $\mathcal{A}$ on the sampled networks. If the result of the algorithm on the original network does not deviate significantly from the values in $\mathbf{X}$, then the result $\mathcal{A}(G)$ is not surprising and its significance is small. Assuming that the sampled datasets are independent and that k is large enough so that $\mathbf{X}$ gives an approximation of the real distribution, then the *empirical p-value* of $X_0 = \mathcal{A}(G)$ is

$$\frac{1}{k+1}\left(\min\{|\{t \mid X_t < X_0\}|, |\{t \mid X_t > X_0\}|\} + 1\right),$$

[7] It would not be difficult to follow the size distribution, considering bigger sets too. But the 5-nodes sets are typical of the mining result, and if the density conditions do not hold for these smaller sized sets they can not hold for bigger sets (if a 6-nodes set is internally dense, it contains 5-nodes sets that are also dense).

i.e., the fraction of the random datasets in which we see a value more extreme than the value in the real data. Another measure for quantifying the significance of the value X_0 is captured by the Z-score $Z = |X_0 - \widehat{X}|/\widehat{\sigma}$, where $\widehat{X} = \mathbf{E}[X_1, \ldots, X_k]$ is the empirical mean of the set $\mathbf{X}$ and $\widehat{\sigma}^2 = \mathbf{Var}[X_1, \ldots, X_k]$ is the empirical variance. Large values of Z indicate that X_0 deviates a lot from the mean of the results obtained on the random datasets.

Significance of our experiments

As explained above, in our experimental evaluation, we validate the statistical significance of our findings using the method of *swap randomization*. We report the results for the networks Flickr-E and Jaiku in Table 3.4.

In all the reasonable settings of the parameters s_0 and σ_0, we tried 10 randomization experiments following the methodology described in [97] In almost all cases, *no core at all was found*. Only in two cases, for the Jaiku network, a very small number of cores appears by chance, still not bringing any community structure (see Table 3.4).

In order to find some cores in the swap randomization version of Flickr-E we had to lower the two parameters s_0 and σ_0. Also in those cases where we can find some cores, δ_{LL} is very low, even lower than δ_{FF}.

The conclusion we draw from the method of swap randomization is that our findings about cores and about their densities are significant, and cannot be explained only by the degree distributions of the nodes in the network.

Table 3.4: Results of swap randomization.

Flickr-E

s_0	σ_0	# cores	$\lvert\mathbb{F}\rvert$	$\lvert\mathbb{L}\rvert$	$\lvert\mathbb{F}\cap\mathbb{L}\rvert$	avg δ_{FF}	avg δ_{LL}
4	30	2 624.2	1161.8	335.6	13.8	0.73	0.18
5	20	30 393.8	1312.1	844.2	26.2	0.78	0.15
5	25	59.7	251.1	72.6	1.1	0.73	0.22
5	30	0	-	-	-	-	-

Jaiku

s_0	σ_0	# cores	$\lvert\mathbb{F}\rvert$	$\lvert\mathbb{L}\rvert$	$\lvert\mathbb{F}\cap\mathbb{L}\rvert$	avg δ_{FF}	avg δ_{LL}
3	30	4.7	14.9	6.7	0	0.007	0.187
3	50	0.2	24.8	1.1	0	0.006	0.044
5	30	0	-	-	-	-	-
5	50	0	-	-	-	-	-

3.4 Experimental Methodology

3.4.1 Core similarity graph

After discovering all the cores in a network, we would like to leverage the output of the mining phase and reason about the community structure of the network. Our first observation is that our mining algorithms discover too many cores, and many of them are very similar (large overlap of leaders and follower). The situation of producing a large number of patterns, which are very similar to each other, is very common in frequent-itemset mining, and there is a significant amount of work dealing with reducing the number of discovered patterns, e.g., see [6, 214, 225, 226].

In this section we focus on the problem of making a concise representation of all the cores discovered in the mining process, and we discuss our algorithm for merging the cores into communities. Even though the cores are found by frequent itemset mining, they have special structure: both the transactions and the items are elements of the same universe and they are all connected to each other with edges in the network structure. Thus, to take advantage of the special structure of our problem we need to go beyond the existing techniques of summarizing frequent itemsets, and we need to design our own methodology for coalescing cores.

Arguably, the leaders of a core (and eventually the leaders of a community) are more important than the followers, since it is them that characterize better the cores and the communities. Thus, we decide to give special emphasis in the role of the leaders. In our coalescing algorithm we merge cores into communities using a similarity measure between cores. We explore three different similarity measures:

(set similarity): two cores are similar if the sets of their leaders are similar;

(link affinity between leaders): two cores are similar if the leaders of one core link to the leaders of the other, and *vice versa*;

(link affinity from followers to leaders): two cores are similar if the followers of one core link to the leaders of the other, and *vice versa*.

We note that, in theory, the two requirements above, may not necessarily occur simultaneously. For instance, a set of nodes with no internal links, compared to itself, has set similarity equal to~1, and link affinity equal to 0. And inversely, two disjoint sets that link fully to each other have set similarity equal to 0, and link affinity equal to 1. And inversely, two disjoint sets that link fully to each other have set similarity equal to 0, and link affinity equal to 1. However, our discovery regarding dense nuclei of cores, implies that in real datasets set similarity is correlated with link affinity.

In a nutshell, our method for coalescing cores into communities is the following.

1. We start with all cores discovered in the mining phase, and we build a *core-similarity graph*, in which cores are nodes and there is an edge between two cores if set similarity of the leaders of the cores is greater than a threshold. Each edge in the core similarity graph is labeled with the corresponding similarity value.

2. We cluster the core/similarity graph by means of agglomerative hierarchical clustering algorithm. The result of the clustering algorithm is a clustering hierarchy tree, in which leaf nodes correspond to cores, and internal nodes correspond to communities obtained by merging all cores in the subtree.

3. Using the clustering hierarchy tree and a user-defined cut-off value we select a cut of the tree and we report the corresponding clustering. Each cluster is a set of cores, and it represents a community. Leaders and followers in the original endorsement network may belong in more than one communities.

We now describe each of the steps in more detail.

Similarity measures Given two sets A and B, a well-known measure of similarity between the two sets is the Jaccard coefficient of the sets: $sim(A, B) = J(A, B) = \frac{|A \cap B|}{|A \cup B|}$. Given two cores $C_1 = (L_1, F_1)$ and $C_2 = (L_2, F_2)$, we define the *set similarity* between the cores to be the Jaccard coefficient of the leader sets, that is $sim_S(C_1, C_2) = J(L_1, L_2)$. For the measuring the link affinity between the leaders of two cores, we use edge density. Given two cores $C_1 = (L_1, F_1)$ and $C_2 = (L_2, F_2)$, we define the *link affinity* between the leaders of the cores to be

$$sim_{LL}(C_1, C_2) = \frac{E(L_1, L_2) + E(L_1, L_2)}{2\,|L_1|\,|L_2|},$$

where $E(L_1, L_2)$ is the number of edges from L_1 to L_2, i.e., $E(L_1, L_2) = |\{(u, v) \mid u \in L_1 \text{ and } v \in L_2\}|$.

For the measuring the link affinity from followers to leaders, we use a similar measure, namely,

$$sim_{FL}(C_1, C_2) = \frac{1}{2}\left(\frac{E(F_1, L_2)}{|F_1|\,|L_2|} + \frac{E(F_2, L_1)}{|F_2|\,|L_1|}\right).$$

Building the core-similarity graph The first step of our algorithm is to build the *core-similarity graph*, in which an edge between two cores C_1 and C_2 declares set similarity $sim_S(C_1, C_2)$ of the cores is greater than a threshold φ_0. Building such a graph is a an instance of the all-pair near-neighbor problem. For the Jaccard similarity measure sim, we can solve this problem using directly the algorithm of Bayardo et al. [31], which is a scalable algorithm to produce all pairs of points with similarity greater than a threshold. The other two measures, sim_{LL} and sim_{FL}, capture edge density between sets of nodes, and thus it is not straightforward how to solve the all-pair near-neighbor problem for these measures. However, as we will show now, the edge-density similarity between two sets of nodes can be expressed as an inner product of appropriately defined vectors. Once we can write our similarity measure as inner product, then we can apply the all-pair near-neighbour algorithm of Bayardo et al. [31]. We demonstrate our reduction for the sim_{LL} measure. The case of sim_{FL} is similar, and we omit the details for brevity.

The main idea is to consider a vector space $\mathbf{R}^{|V|}$, that is, one dimension for each node in the network. Now, for each core $C = (L, F)$ we define two vectors in $\mathbf{R}^{|V|}$

associated to it, a "membership" vector $\mathbf{x}_C$ and an "in-link" vector $\mathbf{y}_C$. In particular the u-th coordinates $\mathbf{x}_C(u)$ and $\mathbf{y}_C(u)$ (corresponding to node u) are defined as

$$\mathbf{x}_C(u) = \begin{cases} \frac{1}{|L|} & \text{if } u \in L, \\ 0 & \text{otherwise,} \end{cases} ,$$

$$\mathbf{y}_C(u) = \frac{|\{(u,v) \mid v \in L\}|}{|L|},$$

that is, $\mathbf{y}_C(u)$ is equal to the number of nodes in the leader set of C that are pointed by u. We can now express the similarity measure sim_{LL} as a function of the vectors $\mathbf{x}_C$ and $\mathbf{y}_C$, i.e.,

$$sim_{LL}(C_1, C_2) = \frac{1}{2}(\mathbf{x}_{C_1} \cdot \mathbf{y}_{C_2} + \mathbf{x}_{C_2} \cdot \mathbf{y}_{C_1}),$$

where $\cdot$ denotes the vector inner product operation.

The last problem that we have to overcome, is that the algorithm of Bayardo et al. is applied to measures expressed as one inner product, and not an average of two inner products, as our measure s_L above. To address this problem, we observe that for an average of two values to be larger than a threshold, at least one of the value has to be larger than the threshold. Thus we compute both scores $\mathbf{x}_{C_1} \cdot \mathbf{y}_{C_2}$ and $\mathbf{x}_{C_2} \cdot \mathbf{y}_{C_1}$ using the algorithm of Bayardo et al., and if one of them is greater than φ_0, we compute the other directly, and we report the pair (C_1, C_2) if the average is greater than φ_0.

3.4.2 Clustering the core-similarity graph

In the next phase we cluster the core-similarity graph using a hierarchical clustering method, introduced in [228] (and available in the CLUTO toolkit[8]). We employ the option of CLUTO for clustering graphs, locally maximizing the traditional criterion function UPGMA. The result of the clustering algorithm is a clustering hierarchy tree, in which leaf nodes correspond to cores in the original core-similarity graph, and internal nodes correspond to communities obtained by merging all cores in the subtree.

3.4.3 Selecting the final clustering

The last step of the algorithm is to use the clustering hierarchy tree produced in the clustering phase in order to produce the final clustering of cores. We assume that we are given the threshold δ_0 on the internal density of communities and we want to produce a clustering in which all the clusters have internal density greater than δ_0. We first observe that the internal density is not monotone with respect to the inclusion in the clustering hierarchy tree: merging two cores may produce communities of either larger or smaller densities.

When clustering data, typically a small number of clusters is preferred since it makes it easier to explain and understand the data. To combine the two desiderata – internally dense communities and small number of clusters – we consider the following approach:

[8] http://glaros.dtc.umn.edu/gkhome/views/cluto

we start from the root of the hierarchy tree and we move downwards the leafs. Once we find an internal node that corresponds to a community with density greater than δ_0 we report this community and we do not consider any further community below that node. We proceed until we produce a cut on the hierarchy tree.

A good choice of the density threshold δ_0 is application dependent, and deciding optimal thresholds for clustering is an elusive objective of the data mining community. If the data analyst has a good idea of the target internal densities in a given application domain, such values can be used. Alternatively, the density threshold controls the number of clusters in the final solution: the higher the threshold the more the number of clusters. Thus, one can adjust the threshold in order to obtain a target number of clusters, if applicable.

3.4.4 Merging cores

When we merge two cores into a community, we have to decide who will be the leaders and who the followers in the resulting community. A natural choice is to consider the union of the leaders and the followers of the two cores. Since being a leader is a stronger property than being a follower, we consider that if a node is a leader for one core, it is a leader for the community, independent of whether or not the node is a follower in the other community. So, if two cores $C_1 = (L_1, F_1)$ and $C_2 = (L_2, F_2)$ needed to be merged in the community $C = (L, F)$, we have

$$L = L_1 \cup L_2, \text{ and } F = (F_1 \cup F_2) \setminus L.$$

We follow the same definition when merging two communities into a larger community.

3.5 Experimental Results

We use the algorithm described in the previous section in order to find communities in two of our endorsement networks, FLICKR-E and JAIKU. For the process of mining cores and clustering we use the following parameters.

For FLICKR-E we mine for all cores with size threshold $s_0 = 4$ leaders and support threshold $\sigma_0 = 150$ followers. The mining process yields 5 777 cores. We then form the core-similarity graphs with similarity threshold $t_0 = 0.2$. The resulting graph for sim_S has maximum degree 1 926 and average degree 582.9. For sim_{LL}, the maximum degree is 4 599 and the average degree 3 621.3. And for sim_{FL}, the maximum degree is 4 599 and the average degree 3 733.6. We then run the CLUTO algorithm on the core-similarity graph to obtain a clustering hierarchy tree, from which we obtain clusterings into communities for various values of the internal density threshold δ_0.

For JAIKU we use size threshold of $s_0 = 3$ leaders and support threshold of $\sigma_0 = 30$ followers, to get a total of 15 132 cores. We create the core-similarity graphs using a similarity threshold of $t_0 = 0.5$. For sim_S the resulting graph has maximum degree 1 711 and average degree 239.7. For sim_{LL}, a maximum degree of 14 401 and an average degree 13 154.1. And for sim_{FL}, a maximum degree of 14 106 and an average degree 12 663.8.

The results of the clustering algorithm, for the two datasets and for a few indicative values of the internal density threshold δ_0, are shown in Table 3.5. We select the value of the threshold δ_0 to get clusterings with number of clusters in the range 10-50. For each value of δ_0, we report indicative statistics including the number of clusters obtained, the total number of leaders and followers, and the average number of leaders and followers in the clusters. As we mentioned before, our community-detection algorithm allows overlapping, a node may be leader or follower in more than one communities, or it may be leader in one and follower in the other. The statistic O_L measures the overlap, as "average membership" of a leader: selecting a leader at random, in how many communities is it a leader? Similarly the statistic O_F measures the average membership of a follower.

We also present all possible combinations of density measures among leaders and followers for intra-cluster and inter-cluster cases. For instance, the inter-cluster density δ_{FL} expresses the probability that there is a link from a follower u to a leader v when u and v belong to different communities. The other intra-cluster and inter-cluster density measures, δ_{LL}, δ_{FL}, δ_{FF} and δ_{LF} are defined in a similar way.

All in all, the quality of the clusterings we obtain is superb: The initial set of leaders and followers that appear in the mined cores is partitioned in relatively small sets of communities, with high intra-cluster leader-leader and follower-leader densities and low inter-cluster leader-leader and follower-leader densities. The density of links among followers and the density of links from leaders to followers is, as expected, much lower, but still the intra-cluster values are significantly higher than the corresponding inter-cluster.

With respect to the similarity measures, we observe that the follower-leader link-based similarity performs the best. It gives clustering with less clusters, higher intra-cluster and lower inter-cluster densities, meaning that the resulting clusters are more compact than the clusters resulting from the other two similarity measures.

3.6 Discussion

In this investigation we have studied the community structure of endorsement (directed) networks, with particular emphasis on the relationship between followers and leaders, and the density of the links among the leaders.

While the problem of community detection is well studied [184], most of the proposed methods in the literature focus on undirected social networks. Moreover the goal of community detection algorithms is usually to optimize measures representing how well the edges are separated in the various clusters; the most commonly used is *modularity* [169, 171].

It is clear that the direction of the links in endorsement networks is an important information that should not be disregarded. The action of a user u favoring a photo of user a v is qualitatively different from reverse action, and it is also different from u and v being friends. Algorithms for community detection in directed social networks are relatively new [108, 142], and the most common practice to deal with directed networks has been to ignore directionality and apply the methods developed for undirected networks. The same consideration holds for bipartite networks: they are simply projected into unipartite

Table 3.5: Clustering results for the two endorsement networks, FLICKR-E and JAIKU, and for the three similarities defined, sim_S, sim_{LL} and sim_{FL}. δ_0: internal density cut-off threshold; $|C|$: number of clusters; $|L|$: total number of leaders; $|F|$: total number of followers; avg$|L_i|$: average number of leaders in the clusters; avg$|F_i|$: average number of followers in the clusters; O_L: overlap of leaders; O_F: overlap of followers; δ_{LL}: leader-leader density; δ_{FL}: follower-leader density; δ_{FF}: follower-follower density; δ_{LF}: leader-follower density.

FLICKR-E

								intra-cluster				inter-cluster			
δ_0	$\lvert C\rvert$	$\lvert L\rvert$	$\lvert F\rvert$	avg$\lvert L_i\rvert$	avg$\lvert F_i\rvert$	O_L	O_F	δ_{LL}	δ_{FL}	δ_{FF}	δ_{LF}	δ_{LL}	δ_{FL}	δ_{FF}	δ_{LF}
Jaccard similarity (sim)															
0.30	12	198	4 815	17.8	480.6	1.0	1.2	0.455	0.322	0.065	0.157	0.020	0.005	0.002	0.003
0.35	22	198	4 818	14.9	399.5	1.6	1.8	0.457	0.326	0.067	0.159	0.024	0.012	0.004	0.008
0.40	30	198	4 822	12.8	345.1	1.9	2.1	0.723	0.485	0.095	0.236	0.022	0.012	0.004	0.008
Leader-leader link similarity (sim_{LL})															
0.30	17	198	4 802	14.7	401.3	1.3	1.4	0.438	0.248	0.048	0.123	0.002	0.004	0.001	0.0003
0.35	26	198	4 805	10.8	309.5	1.4	1.7	0.479	0.272	0.053	0.135	0.002	0.005	0.001	0.0005
0.40	33	198	4 805	9.1	267.3	1.5	1.8	0.539	0.311	0.061	0.155	0.002	0.005	0.001	0.0006
Follower-leader link similarity (sim_{FL})															
0.30	12	198	4 808	18.0	478.3	1.1	1.2	0.618	0.366	0.064	0.179	0.005	0.003	0.001	0.0008
0.35	15	198	4 806	14.9	401.7	1.1	1.2	0.619	0.367	0.065	0.180	0.005	0.004	0.001	0.0009
0.40	17	198	4 808	13.2	365.1	1.1	1.3	0.731	0.445	0.081	0.218	0.005	0.004	0.001	0.0009

JAIKU

								intra-cluster				inter-cluster			
δ_0	$\lvert C\rvert$	$\lvert L\rvert$	$\lvert F\rvert$	avg$\lvert L_i\rvert$	avg$\lvert F_i\rvert$	O_L	O_F	δ_{LL}	δ_{FL}	δ_{FF}	δ_{LF}	δ_{LL}	δ_{FL}	δ_{FF}	δ_{LF}
Jaccard similarity (sim)															
0.10	5	310	2059	77.8	466.2	1.25	1.13	0.081	0.055	0.012	0.013	0.005	0.002	0.001	0.0004
0.15	17	310	2084	25.35	156.47	1.39	1.28	0.140	0.110	0.027	0.027	0.009	0.004	0.002	0.001
0.20	21	310	2084	20.57	127.28	1.39	1.28	0.202	0.154	0.037	0.051	0.008	0.004	0.002	0.001
Leader-leader link similarity (sim_{LL})															
0.10	47	310	2041	9.2	72.6	1.390	1.671	0.219	0.128	0.020	0.029	0.002	0.005	0.001	0.0007
0.15	58	310	2041	7.4	59.1	1.393	1.679	0.316	0.218	0.040	0.050	0.002	0.004	0.001	0.0006
0.20	58	310	2041	7.4	59.1	1.393	1.679	0.316	0.218	0.040	0.050	0.002	0.004	0.001	0.0006
Follower-leader link similarity (sim_{FL})															
0.10	9	310	2037	34.5	230.5	1.003	1.019	0.226	0.130	0.019	0.029	0.0005	0.0002	0.0001	0.00005
0.15	13	310	2037	23.9	160.0	1.003	1.021	0.284	0.172	0.026	0.039	0.0004	0.0002	0.0001	0.00005
0.20	17	310	2037	18.6	125.8	1.022	1.050	0.317	0.203	0.033	0.046	0.0004	0.0005	0.0002	0.00009

networks. This is certainly a loss of relevant information, and it can lead to incorrect results, as it has been recently shown by Guimerà et al. [108].

One of the first papers to address the problem of community detection in bipartite networks is the paper of Guimerà et al. [108]. They consider a bipartite graph with *actors* on one side and *teams* on the other, and they propose to optimize a measure of *bipartite modularity*, which adapts modularity to the bipartite case. They suggest the use of the same approach in directed networks, by first projecting the network onto a bipartite graph. Another recent attempt to address the problem of community discovery in directed networks by means of modularity optimization is proposed by Leicht and Newman [142]. Neither of the two approaches above consider overlapping communities, which is a natural condition in endorsement networks.

The most common approach to detect overlapping communities is the method of *clique percolation* [179]. According to this method, the definition of the communities is based on discovering k-cliques and merging them when they share $k-1$ nodes. This approach has been later extended by the same authors [180] to deal with directed networks by considering *directed k-cliques*, which are complete sub-graphs of size k in which an ordering can be made such that between any pair of nodes there is a directed link from the higher order node towards the lower one.

The latter approach has various similarities to our clustering proposal. However, our method, which is focused on the followers-leaders relationship, uses bicliques instead of directed k-cliques as the basic community footprint. Nevertheless, it is surely interesting to try to apply those methods for community detection to endorsement networks, and clearly the most suitable seems to be the directed clique percolation method. We leave this for future work.

The Web itself can be considered an endorsement network, as it is a directed graph where a link recognizes some form of authoritativeness. Detecting communities of web pages has received attention since the influential work of Kumar et al. [132]. The work of Flake et al. [83], using maximum flow optimization on the web graph to to find topically related communities is also related.

4 Conclusions and Future Research

Summary of the experiment We have studied the relationship between followers and leaders and the induced community structure in endorsement networks, which we defined as directed networks where an edge represents some form of endorsement from a user to another. We applied frequent-pattern mining techniques in order to mine cores, which we define as bicliques of many followers linking to a few leaders. We perceive cores to be the footprints of communities in endorsement social networks. We use real networks to mine cores, and we empirically analyze the cores we mine. We discover that the leaders of the various cores endorse each other, creating a very dense leadership nucleus around which communities are gathered.

We have then developed a novel clustering algorithm to coalesce similar cores into larger communities, while maintaining a high density in their leadership nucleus. Our experiments with coalescing cores into communities show that we are able to discover overlapping communities, in which, on one hand, both followers and leaders link to leaders in their community, but do not link to leaders of other communities.

Lessons learnt Starting from a thorough empirical study of real world networks, we started from the hypothesis that a footprint of community structure in social endorsement networks are bicliques from followers to leaders. In order to find such bicliques in a large endorsement network, we have used highly scalable *frequent itemset mining* techniques on an adjacency-list representation of the network.

Our method produces many similar and redundant cores, which presumably are different footprints of the same community. Thus we have proposed a novel clustering technique in order to coalesce similar cores into meaningful communities. For the clustering algorithm we need to define a measure of similarity between cores. We explore different alternatives that rely on set similarity and on edge density between followers and leaders of the cores. As a technical contribution, we show how to express the edge density measure as an inner product of appropriately defined vectors, and therefore obtain significant computational gains.

Through our analysis of real-world endorsement networks we demonstrate the feasibility and the effectiveness of our approach to discover:

- **Large cores with very dense nuclei:** endorsement networks contains large bicliques from a set of followers to a set of leaders. The set of leaders (nucleus) of a core almost always exhibits an extremely high internal density.
- **Communities:** by coalescing cores we find a reasonably small number of communities having a very large followers base, while still maintaining a very high density in their leadership nucleus.

Our experiments demonstrate the feasibility and effectiveness of our approach: by coalescing cores we find a reasonably small number of communities having a very large followers base, while still maintaining a very high density in their leadership nucleus.

Results Coming back to the original goals of this investigation, we have obtained advances in both our general goals. We have studied the community structure of endorsement social networks, and applied our methods to large scale real-world data. And, using the uncovered community structure, we have defined the notions of *leaders* and *followers*, as a step to analyze the spread of influence through these networks.

About our specific goals, we have:

1. Proposed a method for community detection in endorsement networks. Our method is based on discovering interesting substructures in the graph: *cores.*
    - Our method has been applied to very large networks, using real-world social network site's data. Our method leverages the discovery of local dense substructures of the graph to obtain interesting communities.
    - Our method takes, as a fundamental piece of information, the direction of the edges in the network.
    - The communities we find are naturally overlapping.
    - We have shown that the presence of such core structures is not produced by randomness, and it is therefore statistically significant.
2. We have discovered that the set of leaders in the core, or *nucleus*, almost always exhibits an extremely high internal density. This phenomenon requires further investigation, but we believe the notions of leader and follower are a step in the direction of understanding information spread and influence within endorsement networks.
3. We have seen that the density of the nuclei of the purely friendship-based networks we have worked with is not as high as it is for endorsement networks. This is a very suggestive discovery, but it requires further investigation over more datasets to be validated.

Future work We plan to extend this work in some directions. The first one would be a more thorough evaluation. As we have seen in section 2.5.4, this is in itself a difficult open problem. We plan on exploring the evaluation techniques that are described in that section, and try to adapt them to the particular setting of our problem. For a good evaluation, we will need to gather new and more comprehensive datasets, so significant work has to be done in this direction.

Our work is related to various techniques in the pattern-mining field. We have used standard frequent itemset techniques to mine cores. A future line of research will be to design methods specifically for the problem of core mining, that possibly include the subsequent coalescing phase.

If we think about our directed network as an $n \times n$ binary matrix, where n is the number of nodes, and $(i, j) = 1$ when $i \rightarrow j$ appears in the network, then finding bicliques means finding rectangles of 1s (also called "tiles") where the two dimensions are $\geq s_0$ and σ_0 respectively. One interesting alternative, that we intend to investigate, is to specify only a constraint on the area, instead of one constraint for each of the two dimensions. Such an objective was studied by Geerts et al. [93], who seek to discover large tiles in 0–1 datasets. Geerts et al. also propose another problem that we can reuse our context, namely mining the largest k tiles without specifying the constraint on the area.

Keeping the binary matrix representation in mind, our cores coalescing phase can be thought as finding larger rectangles by relaxing the constraint of having only 1s inside. This is also a well-studied problem, see for instance [227, 34]. In our future investigation we plan to apply those methods in such a way to have a unique mining phase, where we find several rectangles made of almost all 1s with few 0s and which all together give a good coverage of the data.

Our work suggests an interesting novel mining problem. As we are interested in finding cores (bicliques) with high density in the leader part, we might define a frequent-pattern problem with all the constraints: $s(C) \geq s_0$ and $\sigma(C) \geq \sigma_0$ (as in Problem 6) plus the additional constraint $\delta_{\mathrm{LL}}(C) \geq \delta_0$ for a given threshold δ_0. The problem of how to push constraints into the frequent pattern mining computation, in order to reduce the search space and thus the computation, without losing any valid solution was introduced in [174]: many different constraints have been studied in the literature (see [39] for a survey). However, how to push a constraint like $\delta_{\mathrm{LL}}(C) \geq \delta_0$ is an interesting open problem.

Another line of future research is to study the spread of information in the network, in relation to the communities we discover. In this context, we plan on studying the feasibility of simulating spreading processes through the network to evaluate the functions of leader and follower nodes.

Finally, we make use of the inverted indexes approach to compute all pairs similarities above a threshold [31]. This can be easily parallelized. For instance, using the Hadoop framework [150].

Author's publications related to this work

Garrido, G., Bonchi, F., and Gionis, A. 2010. On the high density of leadership nuclei in endorsement social networks. In *Proceedings of the 19th international Conference on World Wide Web* (Raleigh, North Carolina, USA, April 26 - 30, 2010). WWW '10. ACM, New York, NY, 1099-1100.

Bibliography

[1] Facebook statistics. At http://www.facebook.com/press/info.php?statistics. Retrieved in August 2010.

[2] T. B. Achacoso and W. S. Yamamoto. *AY's Neuroanatomy of C. Elegans for Computations*. CRC Press, 1991.

[3] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[4] L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.

[5] P. Adams. The real life social network. Retrieved in August 2010 from www.slideshare.net/padday/the-real-life-social-network-v2. Presentation.

[6] F. Afrati, A. Gionis, and H. Mannila. Approximating a collection of frequent sets. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 12–19, 2004.

[7] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. *Proceedings of the international conference on Web search and web data mining*, pages 183–194, 2008.

[8] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1993.

[9] Y. Ahn, J. Bagrow, and S. Lehmann. Communities and hierarchical organization of links in complex networks. *Arxiv preprint arXiv:0903.3178*, 2009.

[10] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, page 180, 2000.

[11] W. Aiello, F. Chung, and L. Lu. Random evolution in massive graphs. *42nd IEEE symposium on Foundations of Computer Science (FOCS'01)*, page 510, 2001.

[12] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47–97, 2002.

[13] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world wide web. *Nature*, 401(6749):130–131, 1999.

[14] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, Jan 2000.

[15] N. Alon. Spectral techniques in graph algorithms. *LATIN'98: Theoretical Informatics*, pages 206–215, 1998.

[16] U. Alon, M. Surette, N. Barkai, and S. Leibler. Robustness in bacterial chemotaxis. *Nature*, 397(6715):168–171, 1999.

[17] L. Amaral and J. Ottino. Complex systems and networks: challenges and opportunities for chemical and biological engineers. *Chemical engineering science*, 59(8-9):1653–1666, 2004.

[18] L. Amaral, A. Scala, M. Barthélémy, and H. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11149, 2000.

[19] C. Anderson and M. Wolff. The web is dead. long live the internet. Available online at http://www.wired.com/magazine/2010/08/ff_webrip, August 17th 2010.

[20] P. Anderson. More is different. *Science*, 177(4047):393–396, 1972.

[21] A. Arenas, L. Danon, A. Diaz-Guilera, P. Gleiser, and R. Guimerà. Community analysis in social networks. *The European Physical Journal B-Condensed Matter*, 38(2):373–380, 2004.

[22] A. Arenas, A. Fernandez, S. Fortunato, and S. Gomez. Motif-based communities in complex networks. *Journal of Physics A-Mathematical and Theoretical*, 41(22):224001–224001, 2008.

[23] S. Asur and B. Huberman. Predicting the future with social media. *Arxiv preprint arXiv:1003.5699*, 2010.

[24] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. *Proceedings of the 16th international conference on World Wide Web*, page 190, 2007.

[25] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[26] A.-L. Barabási and Z. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[27] M. Barbaro and T. Zeller. Ny times. a face is exposed for aol searcher no. 4417749. Retrieved at http://www.nytimes.com/2006/08/09/technology/09aol.html, August 2006.

[28] A. Barrat and M. Weigt. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

[29] M. Barthélémy and L. Amaral. Small-world networks: Evidence for a crossover picture. *Physical Review Letters*, 82(15):3180–3183, 1999.

[30] R. J. Bayardo. Efficiently mining long patterns from databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'98)*, pages 85–93.

[31] R. J. Bayardo, Y. Ma, and R. Srikant. Scaling up all pairs similarity search. *Proceedings of the 16th international conference on World Wide Web*, pages 131–140, 2007.

[32] D. Beer. Social network (ing) sites... revisiting the story so far: A response to danah boyd & nicole ellison. *Journal of Computer-Mediated Communication*, 13(2):516–529, 2008.

[33] A. Ben-Hur and I. Guyon. *Functional Genomics: Methods and Protocols Humana press*, chapter Detecting stable clusters using principal component analysis, pages 159–182. 2003.

[34] J. Besson, C. Robardet, and J.-F. Boulicaut. Mining a new fault-tolerant pattern type as an alternative to formal concept discovery. In *Proceedings of the 14th International Conference on Conceptual Structures (ICCS 2006).*

[35] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. *Proceedings of the 21st annual international ACM SIGIR Conference*, Jan 1998.

[36] B. Bollobás. The diameter of random graphs. *Transactions of the American Mathematical Society*, 267(1):41–52, 1981.

[37] B. Bollobás and O. Riordan. Clique percolation. *Random Structures and Algorithms*, 2009.

[38] F. Bonchi, F. Giannotti, A. Mazzanti, and D. Pedreschi. Exante: Anticipated data reduction in constrained pattern mining. In *Proceedings of 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'03)*, pages 59–70.

[39] F. Bonchi and C. Lucchese. Extending the state-of-the-art of constraint-based pattern discovery. *Data and Knowledge Engineering (DKE)*, 60(2):377–399, 2007.

[40] S. Borgatti, A. Mehra, D. Brass, and G. Labianca. Network analysis in the social sciences. *Science*, 323(5916):892, 2009.

[41] K. Börner, S. Sanyal, and A. Vespignani. Network science. *Annual Review of Information Science & Technology, Information Today, Inc./American Society for Information Science and Technology, Medford, NJ*, 41:537–607, 2007.

[42] R. Botafogo, E. Rivlin, and B. Shneiderman. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems (TOIS)*, 10(2):142–180, 1992.

[43] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177, 2001.

[44] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. *On modularity-np-completeness and beyond.* 2006.

[45] U. Brandes and T. Erlebach, editors. *Network analysis: methodological foundations.* Number 3418 in Lecture Notes in Computer Science. Springer.

[46] A. Z. Broder. A taxonomy of web search. *ACM SIGIR Forum*, 36(2):3–10, 2002.

[47] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

[48] J. J. Brown and P. H. Reingen. Social ties and word-of-mouth referral behavior. *The Journal of Consumer Research*, 14(3):350–362, 1987.

[49] M. Burke, C. Marlow, and T. Lento. Social network activity and social well-being. *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1909–1912, 2010.

[50] R. S. Burt. *Structural Holes: The Social Structure of Competition.* Harvard University Press, Cambridge, MA, 1995.

[51] R. F. I. Cancho and R. Solé. The small world of human language. *Proceedings: Biological Sciences*, 268(1482):2261–2265, Nov 2001.

[52] K. Carley. A theory of group stability. *American Sociological Review*, 56(3):331–354, 1991.

[53] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences*, 104(27):11150, 2007.

[54] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys (CSUR)*, 38(1), 2006.

[55] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data.* Morgan-Kauffman, 2002.

[56] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. 2002.

[57] H. Champ. Flickr blog - 4,000,000,000. Retrieved in August 2010 from http://blog.flickr.net/en/2009/10/12/4000000000, October 12 2009.

[58] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through url ordering. *Computer Networks and ISDN Systems*, Jan 1998.

[59] A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Arxiv preprint cond-mat/0408187*, 2004.

[60] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences*, 103(7):2015, 2006.

[61] P. Costa. La utilització d'internet per part de barack obama transforma la comunicació política. *Quaderns del CAC*, Jan 2009.

[62] danah m boyd. None of this is real. *in Structures of Participation (ed. Joe Karaganis)*, pages 1–34, Jul 2007.

[63] danah m boyd. Taken out of context: American teen sociality in networked publics. *PhD Dissertation. University of California-Berkeley, School of Information.*, 2008.

[64] danah m boyd and N. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210, 2007.

[65] danah m boyd and J. Heer. Profiles as conversation: Networked identity performance on friendster. *Proceedings of the 39th Hawaii International Conference on System Sciences*, 2006.

[66] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *J. Stat. Mech.*, 2005(09):P09008–P09008, Sep 2005.

[67] G. F. Davis and H. R. Greve. Corporate elite networks and governance changes in the 1980s. *The American Journal of Sociology*, 103(1):1–37, Apr 1997.

[68] I. de Sola Pool and M. Kochen. Contacts and influence. *Social Networks*, Jan 1979.

[69] W. Donath and A. Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973.

[70] S. Dongen. Graph clustering by flow simulation. *University of Utrecht*, 2000.

[71] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. *Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms*, pages 291–299, 1999.

[72] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):27104, 2005.

[73] R. Dunbar. *Grooming, Gossip and the Evolution of Languages*. Harvard University Press, Cambridge, MA, 1998.

[74] D. Easley and J. Kleinberg. Networks, crowds, and markets: Reasoning about a highly connected world [draft version, june 10, 2010]. 2010.

[75] H. Ebel, L. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Phys. Rev. E*, 66(3):35103, 2002.

[76] N. Ellison, C. Steinfield, and C. Lampe. Spatially bounded online social networks and social capital. *International Communication Association*, 36, 2006.

[77] S. Eubank, H. Guclu, V. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[78] T. Evans and R. Lambiotte. Line graphs, link partitions and overlapping communities. *Arxiv preprint arXiv:0903.2181*, 2009.

[79] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, page 262, 1999.

[80] I. Farkas, D. Ábel, G. Palla, and T. Vicsek. Weighted network modules. *New Journal of Physics*, 9(6):180, 2007.

[81] C. Fiduccia and R. Mattheyses. A linear-time heuristic for improving network partitions. *19th Conference on Design Automation, 1982*, pages 175–181, 1982.

[82] G. Flake, S. Lawrence, and C. Giles. Efficient identification of web communities. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–160, 2000.

[83] G. Flake, S. Lawrence, C. Giles, and F. Coetzee. Self-organization and identification of web communities. *Computer*, 35(3):66–70, 2002.

[84] S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

[85] S. Fortunato and C. Castellano. Community structure in graphs. 2007.

[86] S. Fortunato, V. Latora, and M. Marchiori. A method to find community structures based on information centrality. *Physical Review E*, 70(5):56104, 2004.

[87] L. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[88] L. C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science.* Empirical Press, Vancouver, 2004.

[89] E. Garfield. "science citation index"-a new dimension in indexing. *Science*, 144(3619):649–654, 1964.

[90] E. Garfield. Citation analysis as a tool in journal evaluation journals can be ranked by frequency and impact of citations for science policy studies. *Science*, 178(4060):471–479, 1972.

[91] E. Garfield. It's a small world after all. *Essays of an Information Scientist*, 4:299–304, Jul 1979.

[92] D. Garlaschelli, G. Caldarelli, and L. Pietronero. Universal scaling relations in food webs. *Nature*, 423(6936):165–168, 2003.

[93] F. Geerts, B. Goethals, and T. Mielikäinen. Tiling databases. In *Proceedings of 7th International Conference on Discovery Science*, 2004.

[94] D. Gfeller, J. Chappelier, and P. D. L. Rios. Finding instabilities in the community structure of complex networks. *Physical Review E*, 72(5):56135, 2005.

[95] R. Ghosh and K. Lerman. Community detection using a measure of global influence. *Imprint*, 30:10, 2008.

[96] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. *Proceedings of Hypertext'98*, pages 225–234, 1998.

[97] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 167–176, New York, NY, USA, 2006. ACM Press.

[98] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821, 2002.

[99] J. Golbeck. Trust and nuanced profile similarity in online social networks. *ACM Transactions on the Web (TWEB)*, 3(4):12, 2009.

[100] S. Golder, D. Wilkinson, and B. Huberman. Rhythms of social interaction: messaging within a massive online network. *Communities and Technologies 2007*, pages 41–66, 2007.

[101] M. S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, Apr 1973.

[102] S. Gregory. Lnai 5211 - a fast algorithm to find overlapping communities in networks. pages 1–16, Jul 2008.

[103] W. Grossman, S. Schelp, R. Rényi, T. Szemerédi, B. Graham, S. Spencer, S. Pomerance, R. Nathanson, P. Nicolas, and B. Milner. Paul erdos: The master of collaboration. *Algorithmics and Combinatorics*, 14:467–475, 1997.

[104] R. Guimerà and L. Amaral. Cartography of complex networks: modules and universal roles. *J. Stat. Mech*, 2001, 2005.

[105] R. Guimerà and L. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.

[106] R. Guimerà, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in organisations. *Arxiv preprint cond-mat/0211498*, 2002.

[107] R. Guimerà, M. Sales-Pardo, and L. Amaral. Classes of complex networks defined by role-to-role connectivity profiles. *Nature physics*, 3(1):63–69, 2006.

[108] R. Guimerà, M. Sales-Pardo, and L. Amaral. Module identification in bipartite and directed networks. *Physical review. E*, Jan 2007.

[109] K. Hafner. Ny times. researchers yearn to use aol logs, but they hesitate. Retrieved at `www.nytimes.com/2006/08/23/technology/23search.html`, August 23 2006.

[110] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 1970.

[111] D. Haydon, M. Chase-Topping, D. Shaw, L. Matthews, J. Friar, J. Wilesmith, and M. Woolhouse. The construction and analysis of epidemic trees with reference to the 2001 uk foot–and–mouth outbreak. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1511):121, 2003.

[112] G. C. Homans. *The Human Group.* Harcourt, Brace and Company, New York, 1950.

[113] N. V. House. Flickr and public image-sharing: distant closeness and photo exhibition. *CHI'07 extended abstracts on Human factors in computing systems*, page 2722, 2007.

[114] B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1):8, 2009.

[115] C. Hübler, H. Kriegel, K. Borgwardt, and Z. Ghahramani. Metropolis algorithms for representative subgraph sampling. *ICDM '08. Eighth IEEE International Conference on Data Mining*, 2008.

[116] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, 2007.

[117] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[118] R. Kannan and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515, 2004.

[119] G. Karypis. CLUTO - a clustering toolkit. Technical Report #02-017, nov 2003.

[120] G. Karypis and V. Kumar. Metis a software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices version 4.0. *University of Minnesota, Department of Comp*, 1998.

[121] B. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307, 1970.

[122] P. Killworth and H. Bernard. The reversal small-world experiment. *Social Networks*, 1(2):159–192, 1979.

[123] Y. Kim, S. Son, and H. Jeong. Linkrank: Finding communities in directed networks. *Arxiv preprint arXiv:0902.3728*, 2009.

[124] M. Kitsak, L. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. Stanley, and H. Makse. Identifying influential spreaders in complex networks. *Arxiv preprint arXiv:1001.5285*, 2010.

[125] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[126] J. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845, 2000.

[127] J. Kleinberg. The small-world phenomenon: an algorithm perspective. *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, page 170, 2000.

[128] J. Kleinberg. Small-world phenomena and the dynamics of information. *Advances in neural information processing systems 14: proceedings of the 2001 conference*, page 431, 2002.

[129] J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models and methods. *Lecture notes in computer science*, pages 1–17, 1999.

[130] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(88), 2006.

[131] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins, and E. Upfal. The web as a graph. *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–10, 2000.

[132] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks-the International Journal of Computer and Telecommunications Networkin*, 31(11):1481–1494, 1999.

[133] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks–the International Journal of Computer and Telecommunication Networking*, 31(11):1481–1494, 1999.

[134] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.

[135] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80(1):16118, 2009.

[136] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11:033015, 2009.

[137] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*, 78(4):46110, 2008.

[138] P. Lange. Publicly private and privately public: Social networking on youtube. *Journal of Computer-Mediated Communication*, 13(1):361–380, 2007.

[139] A. N. Langville and C. D. Meyer. *Google's Pagerank and Beyond: The Science of Search Engine Rankings.* University Presses of CA, July 2006.

[140] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701, 2001.

[141] E. Leicht and M. Newman. Community structure in directed networks. *Physical Review Letters*, 100(11), 2008.

[142] E. Leicht and M. Newman. Community structure in directed networks. *Physical Review Letters*, 100(11):118703, 2008.

[143] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

[144] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506, 2009.

[145] J. Leskovec and C. Faloutsos. Sampling from large graphs. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 636, 2006.

[146] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. *Proceedings of the 19th international conference on World wide web*, pages 641–650, 2010.

[147] M. Levene and A. Poulovassilis. Web dynamics. *Computer Networks*, 39(3):221–223, 2002.

[148] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623, 2005.

[149] F. Liljeros, C. Edling, L. Amaral, H. Stanley, and Y. Åberg. The web of human sexual contacts. *Nature*, 411(6840):907–908, 2001.

[150] J. Lin. Brute force and indexed approaches to pairwise document similarity comparisons with mapreduce. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*.

[151] R. D. Luce and A. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14:95—116, 1949.

[152] A. Maiya and T. Berger-Wolf. Sampling community structure. *Proceedings of the 19th international conference on World wide web*, pages 701–710, 2010.

[153] C. Marlow, L. Byron, T. Lento, and I. Rosenn. Maintained relationships on facebook. Retrieved in August 2010 at http://overstated.net/2009/03/09/maintained-relationships-on-facebook., 2009.

[154] M. Mathioudakis and N. Koudas. Efficient identification of starters and followers in social media. *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 708–719, 2009.

[155] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: Can we trust what we rt? *1st Workshop on Social Media Analytics (SOMA '10)*, 1060:10.

[156] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, , and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1953.

[157] L. Meyers, B. Pourbohloul, M. Newman, D. Skowronski, and R. Brunham. Network theory and sars: predicting outbreak diversity. *Journal of theoretical biology*, 232(1):71–81, 2005.

[158] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[159] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, Jan 2004.

[160] M. Mizruchi. What do interlocks do? an analysis, critique, and assessment of research on interlocking directorates. *Annu. Rev. Sociol.*, 22:271–298, 1996.

[161] J. Montoya. Small world patterns in food webs. *Journal of theoretical biology*, 214(3):405–412, 2002.

[162] J. Moreno and H. Jennings. Statistics of social configurations. *Sociometry*, pages 342–374, 1938.

[163] J. L. Moreno. *Who Shall Survive?* Beacon House, Beacon, NY, 1953. Available online at www.asgpp.org/docs/WSS/WSS%20Index/WSS%20index.html.

[164] A. Narayanan and V. Shmatikov. Robust de-anonymization of large datasets (how to break anonymity of the netflix prize dataset). *University of Texas at Austin*, 2008.

[165] A. Narayanan and V. Shmatikov. De-anonymizing social networks. *2009 30th IEEE Symposium on Security and Privacy*, pages 173–187, 2009.

[166] M. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(1):16132, 2001.

[167] M. Newman. The structure and function of complex networks. *Arxiv preprint cond-mat/0303516*, 2003.

[168] M. Newman. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):321–330, 2004.

[169] M. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133, 2004.

[170] M. Newman, S. Forrest, and J. Balthrop. Email networks and the spread of computer viruses. *Phys. Rev. E*, 66(3):035101, Sep 2002.

[171] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):26113, 2004.

[172] M. Newman and E. Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23):9564, 2007.

[173] T. news blog CNET News.com. AOL sued over web search data release. Retrieved at http://news.cnet.com/8301-10784_3-6119218-7.html.

[174] R. T. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained associations rules. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD'98)*, 1998.

[175] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri. Extending the definition of modularity to directed graphs with overlapping communities. *J. Stat. Mech*, 2009.

[176] J. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332, 2007.

[177] T. O'Reilly. What is web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies*, (65):17–37, Nov 2007.

[178] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Technical Report. Stanford InfoLab.*, 1998.

[179] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *NATURE-LONDON-*, 435(7043):814, 2005.

[180] G. Palla, I. Farkas, P. Pollner, I. Derenyi, and T. Vicsek. Directed network modules. *New Journal of Physics*, 9(6):186, 2007.

[181] S. Pimm, J. Lawton, and J. Cohen. Food web patterns and their consequences. *Nature*, 350(6320):669–674, 1991.

[182] P. Pollner, G. Palla, and T. Vicsek. Preferential attachment of communities: The same principle, but a higher level. *EPL (Europhysics Letters)*, 73:478, 2006.

[183] P. Pons and M. Latapy. Computing communities in large networks using random walks. *Lecture notes in computer science*, 3733:284, 2005.

[184] M. Porter, J. Onnela, and P. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097, 2009.

[185] A. Potanin, J. Noble, M. Frean, and R. Biddle. Scale-free geometry in oo programs. *Communications of the ACM*, Jan 2005.

[186] A. Pothen. Graph partitioning algorithms with applications to scientific computing. *ICASE LaRC Interdisciplinary Series in Science and Engineering*, 4:323–368, 1997.

[187] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101(9):2658–2663, 2004.

[188] A. Rapoport and W. Horvath. A study of a large sociogram. *Behavioral Science*, 6(4):279–291, 1961.

[189] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):16110, 2006.

[190] J. Reichardt and D. White. Role models for complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 60(2):217–224, 2007.

[191] S. Rice. The identification of blocs in small political bodies. *The American Political Science Review*, 21(3):619–627, 1927.

[192] M. T. Rivera, S. B. Soderstrom, and B. Uzzi. Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *Annu. Rev. Sociol.*, 36(1):91–115, Jun 2010.

[193] M. Rosvall and C. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327, 2007.

[194] M. Rosvall and C. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118, 2008.

[195] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

[196] A. Sala, L. Cao, C. Wilson, R. Zablit, H. Zheng, and B. Zhao. Measurement-calibrated graph models for social network experiments. *Proceedings of the 19th international conference on World wide web*, pages 861–870, 2010.

[197] M. Sales-Pardo, R. Guimerà, A. Moreira, and L. Amaral. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104(39):15224, 2007.

[198] S. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[199] B. Schneier. A taxonomy of social networking data. *IEEE Security & Privacy*, Jan 2010.

[200] M. Sigman and G. Cecchi. Global organization of the wordnet lexicon. *Proceedings of the National Academy of Sciences of the United States of America*, 99(3):1742, 2002.

[201] R. Solé. *Redes Complejas.* Tusquets Editores, Barcelona, January 2009.

[202] R. Solomonoff and A. Rapoport. Connectivity of random nets. *Bulletin of Mathematical Biology*, 13(2):107–117, 1951.

[203] L. Spencer and R. Pahl. *Rethinking Friendship: Hidden Solidarities Today*. Princeton University Press, 2006.

[204] D. Spielman and S. Teng. Spectral partitioning works: Planar graphs and finite element meshes. *Linear Algebra and its Applications*, 421(2-3):284–305, 2007.

[205] O. Sporns, D. Chialvo, M. Kaiser, and C. Hilgetag. Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8(9):418–425, 2004.

[206] R. Stoeckl, P. Rohrmeier, and T. Hess. Motivations to produce user generated content: differences between webloggers and videobloggers. *Proceedings of 20th Bled eConference*, 2007.

[207] F. Stutzman. An evaluation of identity-sharing behavior in social network communities. *International Digital and Media Arts Journal*, 3(1):10–18, 2006.

[208] The Economist. Untangling the social web software. Retrieved in September 2010 from http://www.economist.com/node/16910031, September 2nd 2010.

[209] The Huffington Post. Twitter user statistics revealed. Available online at http://www.huffingtonpost.com/twitter-user-statistics-r_n_537992.html. Retrieved in August 2009, April 14th 2010.

[210] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[211] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *International AAAI Conference on Weblogs and Social Media, Washington, DC*, 2010.

[212] J. Tyler, D. Wilkinson, and B. Huberman. E-mail as spectroscopy: Automated discovery of community structure within organizations. *The Information Society*, 21(2):143–153, 2005.

[213] K. Wakita and T. Tsurumi. Finding community structure in mega-scale social networks. *Proceedings of the 16th international conference on World Wide Web*, page 1276, 2007.

[214] C. Wang and S. Parthasarathy. Summarizing itemset patterns using probabilistic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

[215] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, November 25 1994.

[216] D. J. Watts. Networks, dynamics, and the small-world phenomenon. *The American Journal of Sociology*, 105(2):493–527, May 1999.

[217] D. J. Watts. The "new" science of networks. 2004.

[218] D. J. Watts, P. Dodds, and M. Newman. Identity and search in social networks. *Science*, 296(5571):1302, 2002.

[219] D. J. Watts and S. Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442, 1998.

[220] R. Weiss and E. Jacobson. A method for the analysis of the structure of complex organizations. *American Sociological Review*, pages 661–668, 1955.

[221] J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270, 2010.

[222] C. Williams and G. Gulati. What is a social network worth? facebook and vote share in the 2008 presidential primaries. *Annual Meeting of the American Political Science Association*, pages 1–17, 2008.

[223] R. Williams and N. Martinez. Simple rules yield complex food webs. *Nature*, 404(6774):180–183, 2000.

[224] F. Wu and B. Huberman. Finding communities in linear time: a physics approach. *The European Physical Journal B-Condensed Matter*, 38(2):331–338, 2004.

[225] D. Xin, J. Han, X. Yan, and H. Cheng. Mining compressed frequent-pattern sets. In *Proceedings of the 31st international conference on Very large data bases (VLDB)*, 2005.

[226] X. Yan, H. Cheng, J. Han, and D. Xin. Summarizing itemset patterns: a profile-based approach. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005.

[227] C. Yang, U. M. Fayyad, and P. S. Bradley. Efficient discovery of error-tolerant frequent itemsets in high dimensions. In *Proceedings of the 7th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'01)*, pages 194–203, 2001.

[228] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the CIKM*, 2002.

[229] H. Zhou. Network landscape from a brownian particle's perspective. *Phys. Rev. E*, 67(4):41908, 2003.

[230] H. Zhou and R. Lipowsky. Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. *Computational Science-ICCS 2004*, pages 1062–1069, 2004.

[231] T. Zhou, J. Liu, and B.-H. Wang. Notes on the algorithm for calculating betweenness. *Chinese Physics Letters*, 23:2327, 2006.

List of Tables

List of Figures