Group Formation in Large Social Networks: Membership, Growth, and Evolution

Lars Backstrom
Dept. of Computer Science
Cornell University, Ithaca NY
lars@cs.cornell.edu

Dan Huttenlocher
Dept. of Computer Science
and Johnson Graduate School
of Management
Cornell University, Ithaca NY
dph@cs.cornell.edu

Jon Kleinberg
Dept. of Computer Science
Cornell University, Ithaca NY
kleinber@cs.cornell.edu

Xiangyang Lan
Dept. of Computer Science
Cornell University, Ithaca NY
xylan@cs.cornell.edu

ABSTRACT

The processes by which communities come together, attract new members, and develop over time is a central research issue in the social sciences — political movements, professional organizations, and religious denominations all provide fundamental examples of such communities. In the digital domain, on-line groups are becoming increasingly prominent due to the growth of community and social networking sites such as MySpace and LiveJournal. However, the challenge of collecting and analyzing large-scale time-resolved data on social groups and communities has left most basic questions about the evolution of such groups largely unresolved: what are the structural features that influence whether individuals will join communities, which communities will grow rapidly, and how do the overlaps among pairs of communities change over time?

Here we address these questions using two large sources of data: friendship links and community membership on LiveJournal, and co-authorship and conference publications in DBLP. Both of these datasets provide explicit user-defined communities, where conferences serve as proxies for communities in DBLP. We study how the evolution of these communities relates to properties such as the structure of the underlying social networks. We find that the propensity of individuals to join communities, and of communities to grow rapidly, depends in subtle ways on the underlying network structure. For example, the tendency of an individual to join a community is influenced not just by the number of friends he or she has within the community, but also crucially by how those friends are connected to one another. We use decision-tree techniques to identify the most significant structural determinants of these properties. We also develop a novel methodology for measuring movement of individuals between communities, and show how such movements are closely aligned with changes in the topics of interest within the communities.

Categories and Subject Descriptors: H.2.8 Database Management: Database Applications – Data Mining

General Terms: Measurement, Theory

Keywords: social networks, on-line communities, diffusion of innovations

This work has been supported in part by NSF grants CCF-0325453, IIS-0329064, CNS-0403340, BCS-0537606, and 0121175, by the Institute for the Social Sciences at Cornell, and by the John D. and Catherine T. MacArthur Foundation.

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

1. INTRODUCTION

The tendency of people to come together and form groups is inherent in the structure of society; and the ways in which such groups take shape and evolve over time is a theme that runs through large parts of social science research [9]. The study of groups and communities is also fundamental in the mining and analysis of phenomena based on sociological data — for example, the evolution of informal close-knit groups within a large organization can provide insight into the organization's global decision-making behavior; the dynamics of certain subpopulations susceptible to a disease can be crucial in tracking the early stages of an epidemic; and the discussions within an Internet-based forum can be used to follow the emergence and popularity of new ideas and technologies. The digital domain has seen a significant growth in the scale and richness of on-line communities and social media, through the rise of social networking sites beginning with Friendster and its relatives, and continuing to more recent systems including MySpace, Facebook, and LiveJournal, as well as media-sharing sites such as Flickr.

Understanding the structure and dynamics of social groups is a natural goal for network analysis, since such groups tend to be embedded within larger social network structures. That is, given a collection of individuals linked in an underlying social network, the groups and communities that they identify with can be thought of as corresponding to subgraphs of this network, growing and overlapping one another in a potentially complex fashion. A group that grows mainly through the agressive recruitment of friends by other friends would appear as a subgraph branching out rapidly over time along links in the network; a group in which the decision to join depends relatively little on the influence of friends might appear in-

stead as a collection of small disconnected components that grows in a "speckled" fashion.[1]

While abstract descriptions such as this — of groups growing concurrently and organically in a large network — are clearly suggestive, the fact is that it has been very hard to make concrete empirical statements about these types of processes. Much of the challenge arises from the difficulty in identifying and working with appropriate datasets: one needs a large, realistic social network containing a significant collection of explicitly identified groups, and with sufficient time-resolution that one can track their growth and evolution at the level of individual nodes. A further challenge has been the lack of a reasonable vocabulary for talking about group evolution — with each group growing in its own particular part of the network, how do we abstract and quantify the common types of patterns that we observe?

The present work: Analyzing Group Formation and Evolution. In this paper we seek to address these challenges, exploring the principles by which groups develop and evolve in large-scale social networks. We consider a number of broad principles about the formation of social groups, concerning the ways in which they grow and evolve, and we formalize concrete questions around them that can be tested on network data.

To do this, we take advantage of rich datasets and computational models for describing the process of group formation. In particular, as our primary sources of data, we make use of two networks that combine the desirable features outlined above: LiveJournal, a social networking and blogging site with several million members and a large collection of explicit user-defined communities; and DBLP, a publication database with several hundred thousand authors over several decades, and where conferences serve as proxies for communities. We will say more about these datasets below; for now, we note the crucial point that we are focusing on networks where the members have *explicitly* identified themselves as belonging to particular groups or communities — we are thus not seeking to solve the unsupervised graph clustering problem of inferring "community structures" in a network (e.g., [14, 15, 16, 20, 28]), since for us the relevant communities have been identified by the members themselves.

We consider three main types of questions.

- **Membership.** What are the structural features that influence whether a given individual will join a particular group?
- **Growth.** What are the structural features that influence whether a given group will grow significantly (i.e. gain a large net number of new members) over time?
- **Change.** A given group generally exists for one or more purposes at any point in time; in our datasets, for example, groups are focused on particular "topics of interest." How do such foci change over time, and how are these changes correlated with changes in the underlying set of group members?

The question of membership is closely related to the well-studied topic of *diffusion of innovation* in the social sciences (see e.g. [31, 33, 34] as well as [13, 21, 30] for more recent applications in the data mining literature). That is, if we view the act of joining a particular group as a kind of behavior that "spreads" through the network, then how does one's probability p of joining a group depend on the friends that one already has in the group? Perhaps the most basic such question is how the probability p depends on the number of friends k that one already has in the group. This is a fundamental question in research on diffusion in social networks, and most mathematical models of this process implicitly posit a model for the dependence of p on k (see e.g. [13, 21, 34]); however, it has to date been easier to explore such models theoretically than to obtain reasonable estimates for them empirically on large-scale data. Here we find that this dependence is remarkably similar for groups in the LiveJournal and DBLP datasets, despite the very different meaning of the groups in these two domains; the probability p increases, but sublinearly so, in the number of friends k belonging to the group. The data suggest a "law of diminishing returns" at work, where having additional friends in a group has successively smaller effect but nonetheless continues to increase the chance of joining over a fixed time window. In the context of diffusion models this result is somewhat surprising, in that it does not appear to be explained well by models that posit logistic or "critical mass" behavior for p versus k.

Beyond this, however, the available data makes possible a much broader investigation of membership in groups. While theoretical models of diffusion have focused primarily on just the effect of k, the number of friends one already has in a group, we would like to understand more generally the structural properties that are most influential in determining membership. Here we do this by applying a decision-tree approach to the question, incorporating a wide range of *structural features* characterizing the individual's position in the network and the subgraph defining the group, as well as *group features* such as level of activity among members. In the process we find that the probability of joining a group depends in subtle but intuitively natural ways not just on the number of friends one has, but also on the ways in which they are connected to one another.

To take one illustrative example: for moderate values of k, an individual with k friends in a group is significantly more likely to join if these k friends are themselves mutual friends than if they aren't. This example fits naturally with known sociological dichotomies on diffusion, and hence it hints at some of the more qualitative processes at work in the communities we are studying.

We adopt a similar approach to the question of growth: given a group, how well can we estimate whether it will grow by a significant fraction of its current size over a fixed time period? We find that reasonable estimation performance can be obtained based purely on the structural properties of the group as a subgraph in the underlying social network. As with membership, relatively subtle structural features are crucial in distinguishing between groups likely to grow rapidly and those not likely to. Again, to focus on one example, groups with a very large number of triangles (consisting of three mutual friends) grow significantly less quickly overall than groups with relatively few triangles. Overall, then, the framework based on decision trees can be viewed as a way to identify the most "informative" structural and group features influencing the growth and membership processes, with the payoff that the resulting features have natural interpretations in terms of the underlying sociological considerations.

Groups not only grow and attract new members — the very characteristics of a group can change over time. A group A may change its focus of interest to become more like some other group B; it may also change its membership to become more like B. The final set of questions that we investigate addresses issues of change in group membership and interests, as well as the extent to which there is a correlation between these two types of change. For instance do

[1]While such social networks are not themselves directly observable, on-line systems can provide rich data on large networks of interactions that are highly reflective of these underlying social networks. As has become customary in the computer science community, we also refer to these observable networks as social networks, while recognizing that they are only a reflection of the complete picture of social interactions.

changes in membership consistently precede or lag changes in interest? While such questions are extremely natural at a qualitative level, it is highly challenging to turn them into precise quantitative ones, even on data as detailed as we have here. We approach this through a novel methodology based on burst analysis [22]; we identify bursts both in term usage within a group and in its membership. We find that these are aligned in time to a statistically significant extent; furthermore, for CS conference data in DBLP, we present evidence that topics of interest tend to cross between conferences earlier than people do.

Related Work. As discussed above, there is a large body of work on identifying tightly-connected clusters within a given graph (see e.g. [14, 15, 16, 20, 28]). While such clusters are often referred to as "communities", it is important to note that this is a very different type of problem from what we consider here — while this clustering work seeks to infer potential communities in a network based on density of linkage, we start with a network in which the communities of interest have already been explicitly identified and seek to model the mechanisms by which these communities grow and change. Dill et al. [11] study implicitly-defined "communities" of a different sort: For a variety of features (e.g. a particular keyword, a name of a locality, or a ZIP code), they consider the subgraph of the Web consisting of all pages containing this feature. Such communities of Web pages are still quite different from explicitly-identified groups where participants deliberately join, as we study here; moreover, the questions considered in [11] are quite different from our focus here.

The use of on-line social networking sites for data mining applications has been the subject of a number of recent papers; see [1, 26] for two recent examples. These recent papers have focused on different questions, and have not directly exploited the structure of the user-defined communities embedded in these systems. Studies of the relationship between different newsgroups on Usenet [4, 35] has taken advantage of the self-identified nature of these online communities, although again the specific questions are quite different.

As noted earlier, the questions we consider are closely related to the *diffusion of innovations*, a broad area of study in the social sciences [31, 33, 34]; the particular property that is "diffusing" in our work is membership in a given group. The question of how a social network evolves as its members' attributes change has been the subject of recent models by Sarkar and Moore [32] and Holme and Newman [19]; a large-scale empirical analysis of social network evolution in a university setting was recently performed by Kossinets and Watts [23]; and rich models for the evolution of topics over time have recently been proposed by Wang and McCallum [36]. Mathematical models for group evolution and change have been proposed in a number of social science contexts; for an approach to this issue in terms of diffusion models, we refer the reader to the book by Boorman and Levitt [3].

2. COMMUNITY MEMBERSHIP

Before turning to our studies of the processes by which individuals join communities in a social network, we provide some details on the two sources of data, LiveJournal and DBLP. LiveJournal (LJ) is a free on-line community with almost 10 million members; a significant fraction of these members are highly active. (For example, roughly $300,000$ update their content in any given 24-hour period.) LiveJournal allows members to maintain journals, individual and group blogs, and — most importantly for our study here — it allows people to declare which other members are their friends and to which communities they belong. By joining a community,

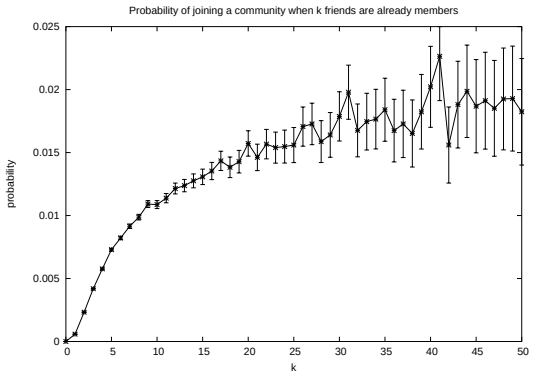

Figure 1: The probability p of joining a LiveJournal community as a function of the number of friends k already in the community. Error bars represent two standard errors.

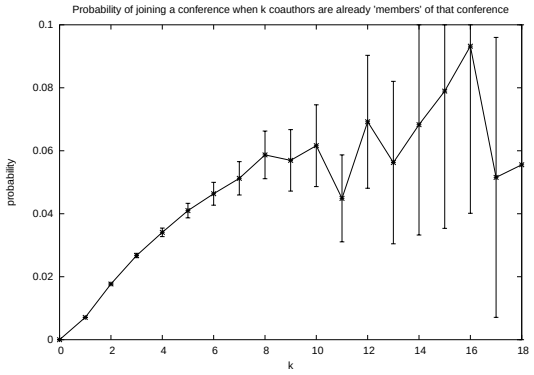

Figure 2: The probability p of joining a DBLP community as a function of the number of friends k already in the community. Error bars represent two standard errors.

one typically gains the right to create new posts in that community and other people's posts become more accessible.

DBLP, our second dataset, is an on-line database of computer science publications, providing the title, author list, and conference of publication for over 400,000 papers. A great deal of work has gone into disambiguation of similar names, so co-authorship relationships are relatively free of name resolution problems. For our purposes, we view DBLP as parallel to the friends-and-communities structure of LiveJournal, with a "friendship" network defined by linking people together who have co-authored a paper, and with conferences serving as communities. We say that a person has joined a community (conference) when he or she first publishes a paper there; and, for this section, we consider the person to belong to the community from this point onward. (See Section 4 for an analysis of changes in community membership that include notions of both joining and leaving.) For simplicity of terminology, we refer to two people in either of LJ or DBLP as "friends" when they are neighbors in the respective networks.

A fundamental question about the evolution of communities is to to determine who will join in the future. As discussed above, if we view membership in a community as a "behavior" that spreads through the network, then we can gain insight into this question from the study of the diffusion of innovation [31, 33, 34].

Table 1: Features.

Feature Set	Feature
Features related to the community, C. (Edges between only members of the community are $E_C \subseteq E$.)	Number of members ($\|C\|$). Number of individuals with a friend in C (the *fringe* of C) . Number of edges with one end in the community and the other in the fringe. Number of edges with both ends in the community, $\|E_C\|$. The number of open triads: $\|\{(u,v,w)\|(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \notin E_C \wedge u \neq w\}\|$. The number of closed triads: $\|\{(u,v,w)\|(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \in E_C\}\|$. The ratio of closed to open triads. The fraction of individuals in the fringe with at least k friends in the community for $2 \leq k \leq 19$. The number of posts and responses made by members of the community. The number of members of the community with at least one post or response. The number of responses per post.
Features related to an individual u and her set S of friends in community C.	Number of friends in community ($\|S\|$). Number of adjacent pairs in S ($\|\{(u,v)\|u,v \in S \wedge (u,v) \in E_C\}\|$). Number of pairs in S connected via a path in E_C. Average distance between friends connected via a path in E_C. Number of community members reachable from S using edges in E_C. Average distance from S to reachable community members using edges in E_C. The number of posts and response made by individuals in S. The number of individuals in S with at least 1 post or response.

2.1 Dependence on number of friends

An underlying premise in diffusion studies is that an individual's probability of adopting a new behavior increases with the number of friends already engaging in the behavior — in this case, the number of friends already in the community.

In Figures 1 and 2 we show this basic relationship for LJ and DBLP respectively: the proportion $P(k)$ of people who join a community as a function of the number k of their friends who are already members. For LJ, this is computed as follows.

- First, we took two snapshots of community membership, roughly one month apart.
- Then we find all triples (u, C, k) such that
  - C is a community, and
  - u is a user who, at the time of the first snapshot, did not belong to C, and
  - u had k friends in C at that time.
- $P(k)$ is then the fraction of such triples (u, C, k) for a given k such that u belonged to C at the time of the second snapshot.

The procedure for DBLP is analogous, except that we use a snapshot for each year, and determine the fraction of individuals who "join" a conference from one year to the next.

The plots for LJ and DBLP exhibit qualitatively similar shapes, dominated by a "diminishing returns" property in which the curve continues increasing, but more and more slowly, even for relatively large numbers of friends k. This forms an interesting contrast to the "S-shaped" curve at the heart of many theoretical models of diffusion, in which the probability of adopting a new behavior follows a logistic function, with slow growth in adoption probability for small numbers of friends k, rapid growth for moderate values of k, and a rapid flattening of the curve beyond this point.

In fact, the curves do exhibit some slight but noticeable "S-shaped" behavior: While the plots mainly show sublinear increase, we observe that they each display a deviation for $k = 0, 1, 2$ — namely, $P(2) > 2P(1)$ for both LJ and DBLP. In other words, the marginal benefit of having a second friend in a community is particularly strong. However the remainder of each plot exhibits diminishing returns as k increases; thus the deviation at $k = 0, 1, 2$ can be seen as a slight "S-shaped" effect before the sublinear behavior takes over. Focusing on the function $P(k)$ for LJ, since the error bars are smaller here, we see that the curve continues increasing even for quite large values of k. Indeed, there is a close fit to a function of the form $P(k) = a \log k + b$ for appropriate a and b.

A key reason that the curve for LJ is quite smooth is that the amount of data used to generate it is very large: the construction of the plot in Figure 1 is based on roughly half a billion triples of the form (u, C, k) with $k > 0$. The analogous number of triples for DBLP is 7.8 million, and the curve becomes noisy at much smaller values of k. This suggests that for computing $P(k)$ as a function of k in the context of diffusion studies, a very large sample may be required to begin seeing the shape of the curve clearly.

We find it striking that the curves for LJ and DBLP have such similar shapes (including the deviations for $k = 0, 1, 2$), given that the types of communities represented by these two datasets have such different characteristics: joining a community is a relatively lightweight operation in LJ, requiring very little investment of effort, whereas the analogous act of joining in the case of the DBLP dataset requires authorship and acceptance of a conference paper.

Curves with a diminishing returns property were also recently observed in independent work of Leskovec et al. [25], in yet another different context — recommendation data for on-line purchases — although the curves in their case become noisier at smaller values of k. The probability of friendship as a function of shared acquaintances and shared classes also exhibits diminishing returns in the work of Kossinets and Watts [23]. It is an interesting question to look for common principles underlying the similar shapes of the curves in these disparate domains.

2.2 A broader range of features

While these curves represent a good start towards membership prediction, they estimate the probability of joining a community based on just a single feature — the number of friends an individual has in the community. We now consider a range of other features related both to the communities themselves and to the topology of the underlying network which could also, in principle, influence the probability of joining a community. By applying decision-tree

techniques to these features we find that we can make significant advances in estimating the probability of an individual joining a community. Table 1 summarizes the features that we use. In addition to features related exclusively to the social network structure, we also generate simple features that serve as indicators of the activity level of a community in LJ (for example, the number of messages posted by members of the community).[2] A recurring principle in our experimental set-up is the following: since our goal is to understand which features from a particular set of structural and activity-based features are most informative, we intentionally control the set of features available to our algorithms. For the strict goal of obtaining high prediction performance, there are other features that could be included that would be less informative for our current purposes.

We now discuss the exact structure of the sets over which we make predictions for both LJ and DBLP.

LiveJournal. For the more detailed studies of membership prediction, we focused on a subset of 875 LJ communities, comparing them from the first LJ snapshot to the second.[3] For the first of these snapshots, we also built the network structure on the communities and their fringes. (We define the *fringe* of a community C to be the set of all non-members of C who have at least one friend in C.) In addition, we collected all posts during the two weeks prior to the initial snapshot. (This two-week period was disjoint from the initial period during which we selected the 875 communities.)

From this information, we created a data point (u, C) for each user u and community C such that u belonged to the fringe of C in the first snapshot. We then estimated the probability each such fringe member would be in the community in the second snapshot. Note that this task is an instance of the general problem of estimating missing values in a matrix: we are given a matrix whose rows correspond to users, whose columns correspond to communities, and whose entries (u, C) indicate whether u joins C in the time interval between the two snapshots. In this way, the set-up is syntactically analogous to what one sees for example in collaborative-filtering-style problems; there too one is attempting to estimate hidden matrix-valued data (e.g. which customers are likely to buy which books). In keeping with our design principle, however, we are interested in performance based only on carefully selected features of the users u and communities C, rather than their actual identities.

We have 17,076,344 data points (u, C), and of these, only 14,488 of represent instances in which user u actually joined community C, for an average rate of 8.48e-4. Note that our task here, to estimate probabilities for individuals joining, is compatible with the low aggregate rate of joining. To make estimates about joining, we grow 20 decision trees. Each of the 875 communities is selected to have all of its fringe members included in the decision tree training set or not with independent probability 0.5. At each node in the decision tree, we examine every possible feature, and every binary split threshold for that feature. Of all such pairs, we select and install the split which produces the largest decrease in entropy [29] (i.e. information gain). We continue to install new splits until there are fewer than 100 positive cases at a node, in which case we install a leaf which predicts the ratio of positives to total cases for that node. Finally, for every case we find the set of decision trees for which that case was not included in the training set used to grow the tree. The average of these predictions gives us a prediction for the case. For the few cases that we include in the training set of every decision tree, we simply predict the baseline 8.48e-4. This technique of model averaging [5] has been shown to be effective in prediction settings such as these.

[2]Due to the much more regimented nature of conference activity, we do not generate analogous activity features for the DBLP dataset.

[3]We chose the 875 communities as follows. We monitored all new posts to all communities during a 10 day period. Of those communities which had at least 1 post, we selected the 700 most active communities along with 300 at random from the others with at least 1 post. For technical reasons, it turned out that we were not able to collect accurate data on the largest of the communities, and hence were forced to discard communities which started with over 1000 members, leaving 875 communities.

Table 2: Prediction performance for single individuals joining communities in LiveJournal. For every individual in the fringe of one of our 875 communities, we estimate the probability that person will join in a one-month interval. We repeat this experiment using 3 sets of features: only the number of friends in the community, features based on post activity (plus basic features: number of friends and community size), and finally the combination of all the features, including the graph-theoretic ones from Table 1.

Features Used	ROCA	APR	CXE
Number of Friends	0.69244	0.00301	0.00934
Post Activity	0.73421	0.00316	0.00934
All	0.75642	0.00380	0.00923

Table 3: Prediction performance for single individuals joining communities in DBLP. For every triple of a year, a conference, and an author who had not published in the conference, but had coauthored with a conference member, we estimate the probability that the author will publish in the conference's next meeting.

Features Used	ROCA	APR	CXE
Number of Friends	0.64560	0.01236	0.06123
All	0.74114	0.02562	0.05808

DBLP. For DBLP we perform a very similar experiment. Here we define the fringe of a conference C in year y to be those people who have not published in C prior to year y, but who have coauthored with at least one person who has published in C prior to y. For every conference, year, and fringe member in that year we create a data point. Of 7,651,013 data points, we find that 71,618 correspond to individuals who join the conference (publish a paper in it) in the year in question. Again, to make predictions we use 20 simple decision trees grown in an identical way to those for LJ.

2.3 Results and Discussion

Table 2 and Table 3 summarize the performance we achieve with these decision trees. For comparison, both tables contain the baseline performance one could achieve by predicting based solely on the number of friends a fringe member already has in the community. In all of our predictions, even the people who are most likely to join a community still have a probability much less than 50%. This makes performance metrics like accuracy meaningless, since if one had to make binary decisions, one would simply predict that no one would join. We thus use performance metrics that are based on the order of predictions: area under the ROC curve (ROCA) and average precision (APR), as well as cross entropy (CXE), which treats predictions as probabilities. The two tables show that we are able to do significantly better by using features beyond the number of friends an individual has in the community.

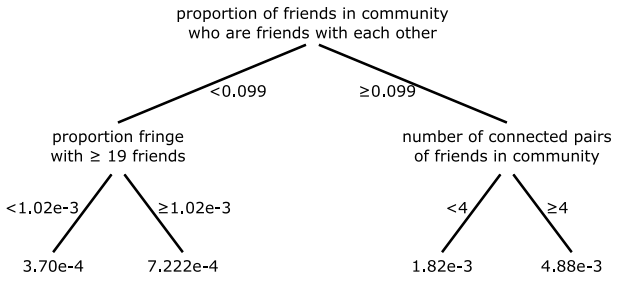

Figure 3: The top two levels of decision tree splits for predicting single individuals joining communities in LiveJournal. The overall rate of joining is 8.48e-4.

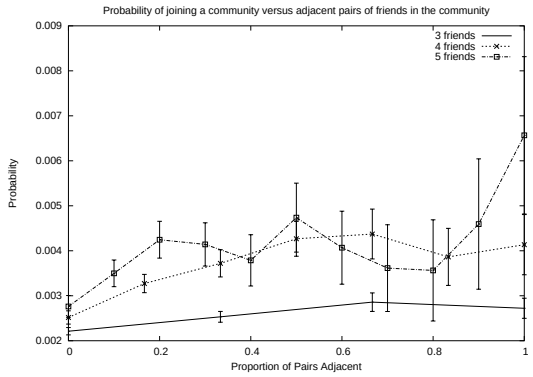

Figure 4: The probability of joining a LiveJournal community as a function of the internal connectedness of friends already in the community. Error bars represent two standard errors.

Internal Connectedness of Friends. The top-level splits in the LJ and DBLP decision trees were quite stable over multiple samples; in Figure 3 we show the top two levels of splits in a representative decision tree for LJ. We now discuss a class of features that proved particularly informative for the LJ dataset: the internal connectedness of an individual's friends.

The general issue underlying this class of feature is the following: given someone with k friends in a community, are they more likely to join the community if many of their friends are directly linked to one another, or if very few of their friends are linked to one another? This distinction turns out to result in a significant effect on the probability of joining. To make this precise, we use the following notation. For an individual u in the fringe of a community, with a set S of friends in the community, let $e(S)$ denote the number of edges with both ends in S. (This is the number of pairs in S who are themselves friends with each other.) Let $\varphi(S) = e(S)/\binom{|S|}{2}$ denote the *fraction* of pairs in S connected by an edge.

We find that individuals whose friends in a community are linked to one another — i.e., those for which $e(S)$ and $\varphi(S)$ are larger — are significantly more likely to join the community. In particular, the top-level decision tree split for the LJ dataset is based on $\varphi(S)$, and in the right branch (when $\varphi(S)$ exceeds a lower bound), the next split is based on $e(S)$. We can see the effect clearly by fixing a number of friends k, and then plotting the joining probability as a function of $\varphi(S)$, over the sub-population of instances where the individual has k friends in the community. Figure 4 shows this relationship for the sub-populations with $k = 3$, 4, and 5; in each case, we see that the joining probability increases as the density of linkage increases among the individual's friends in the community.

It is interesting to consider such a finding from a theoretical perspective — why should the fact that your friends in a community know each other make you more likely to join? There are sociological principles that could potentially support either side of this dichotomy.[4] On the one hand, arguments based on *weak ties* [17] (and see also the notion of *structural holes* in [6]) support the notion that there is an informational advantage to having friends in a community who do not know each other — this provides multiple "independent" ways of potentially deciding to join. On the other hand, arguments based on social capital (e.g. [8, 9]) suggest that there is a trust advantage to having friends in a community who know each other — this indicates that the individual will be supported by a richer local social structure if he or she joins. Thus, one possible conclusion from the trends in Figure 4 is that trust advantages provide a stronger effect than informational advantages in the case of LiveJournal community membership.

The fact that edges among one's friends make community membership more likely is also consistent with observations made in recent work of Centola, Macy, and Eguiluz [7]. They contend that instances of successful social diffusion "typically unfold in highly clustered networks" [7]. In the case of LJ and DBLP communities, for example, Macy observes that links among one's friends may contribute to a "coordination effect," in which one receives a stronger net endorsement of a community if it is a shared focus of interest among a group of interconnected friends [27].

Relation to Mathematical Models of Diffusion. There are a number of theoretical models for the diffusion of a new behavior in a social network, based on simple mechanisms in which the behavior spreads contagiously across edges; see for example [12, 21, 34] for references. Many of these models operate in regimented time steps: at each step, the nodes that have already adopted the behavior may have a given probability of "infecting" their neighbors; or each node may have a given threshold d, and it will adopt the behavior once d of its neighbors have adopted it.

Now, it is an interesting question to consider how these models are related to the empirical data in Figures 1 and 2. The theoretical models posit very simple dynamics by which influence is transmitted: in each time step, each node assesses the states of its neighbors in some fashion, and then takes action based on this information. The spread of a real behavior, of course, is more complicated, and our measurements of LJ and DBLP illustrate this: we observe the behavior of an individual node u's friends in one snapshot, and then u's own behavior in the next, but we do not know (i) when or whether u became aware of these friends' behavior, (ii) how long it took for this awareness to translate into a decision by u to act, and (iii) how long it took u to actually act after making this decision. (Imagine, for example, a scenario in which u decides to join a community after seeing two friends join, but by the time u actually joins, three more of her friends have joined as well.) Moreover, for any given individual in the LJ and DBLP data, we do not know how far along processes (i), (ii), and (iii) are at the time of the first snapshot — that is, we do not know how much of the information contained in the first snapshot was already known to the individual, how much they observed in the interval between the first and second snapshots, and how much they never observed.

These considerations help clarify what the curves in Figures 1 and 2 are telling us. The concrete property they capture is the mea-

[4]We thank David Strang for helping to bring the arguments on each side into focus.

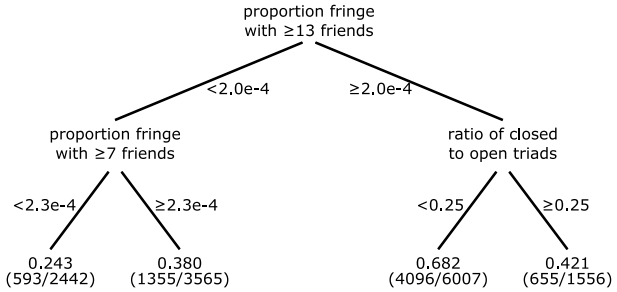

Figure 5: The top two levels of decision tree splits for predicting community growth in LiveJournal.

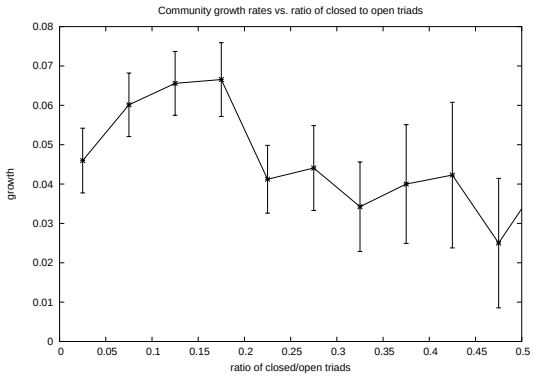

Figure 6: The rate of community growth as a function of the ratio of closed to open triads: having a large density of closed triads (triangles) is negatively related to growth. Error bars represent two standard errors.

sured probability of adoption over a fixed time window, based on observed properties of an earlier snapshot — and they do this for network data on a scale that has been hard to obtain in earlier social science studies of this phenomenon. Building on this, it is a natural challenge to relate the data underlying these curves to more purely operational models by which influence is spread through a network, and potentially to assess whether such models are reasonable approximations of real diffusion processes.

3. COMMUNITY GROWTH

We now turn to a different but related prediction task: identifying which communities will grow significantly over a given period of time. We apply decision tree techniques to this task as well, using the community features given in the first half of Table 1.

For this experiment, our features come from two snapshots of community membership and social network topology, taken roughly 4 months apart. Since the behavior of extremely small communities is determined by many factors that are not observable from the network structure, we perform our experiments only on those communities which had at least 100 members at the time of the first snapshot. We say that a community has a *growth rate* of $x\%$ if its size in the second snapshot is $x\%$ larger than its size in the first snapshot. Over all communities, the mean growth rate was 18.6%, while the median growth rate was 12.7%.

We cast this problem directly as a binary classification problem in which class 0 consists of communities which grew by less than 9%, while class 1 consists of communities which grew by more than 18%. We find that by excluding the middle we achieve more meaningful estimates of performance, as it is unreasonable to expect good performance in the region around a simple threshold. This leaves us a data set with 13570 communities, 49.4% of which are class 1.

To make predictions on this dataset we again use binary decision trees. Because this data set is smaller and more balanced, we install binary splits until a node has less than 50 data points, in which case we install a leaf which predicts the fraction of positive instance at that point. We grow 100 decision trees on 100 independent samples of the full dataset. For a particular test case, we make a prediction for that case using all of the decision trees which were not grown using that case.

3.1 Results

For comparison, we start by considering a number of simple baseline predictions, shown in Table 4. Using the same technique of averaging trees, but with only a single feature, we construct three baselines. The first feature for comparison is simply the size of the community. One might suspect that communities with a large number of people became large for a reason and are likely to continue growing. The second baseline uses the number of people in the fringe of the community, as these are the people most likely to join. Finally, we use the ratio of these two features — the size of the fringe divided by the size of the community — as well as a combination of all three.

Table 4: Results for predicting community growth: baselines based on three different features, and performance using all features.

Features Used	ROCA	APR	CXE	ACC
Fringe	0.55874	0.53560	1.01565	0.54451
Community Size	0.52096	0.52009	1.01220	0.51179
Ratio of Fringe to Size	0.56192	0.56619	1.01113	0.54702
Combination of above 3	0.60133	0.60463	0.98303	0.57178
All Features	0.77070	0.77442	0.82008	0.70035

Table 4 shows that none of these simple features gives good performance by itself. While they each perform slightly better than random guessing, the difference is small. Furthermore, using these three baseline features in combination still does not yield very impressive results: an ROCA of 0.60133 as compared to 0.5 for random guessing.

By including the full set of features described previously, however, we find that we can make predictions with reasonably good performance. ROCA increases to 0.77070 , while accuracy goes up to 70%. Other performance metrics indicate improvement on similar scales. Furthermore, accuracy on the fastest growing communities is as high as 80%.

3.2 Discussion of Results

It is informative to look at which features are being used at the top-level splits made by the decision trees. Figure 5 shows the top 2 splits installed by a representative tree. While the features and splits in the tree varied depending on the sample, the top 2 splits were quite stable, with only minor variations between samples. The first of these is the fraction of people that have a large number of friends in the community. Given the results of the previous section, this is intuitively natural. At the top level, we see that communities having a higher percentage of fringe members with at least 13 friends in the community are much more likely to be of class 1. Furthermore,

of the communities with relatively few such fringe members, the next split is based on the percentage of individuals with 7 friends in the community.

A second class of features, also important for community growth though for less intuitively apparent reasons, is the density of triangles. (See the right subtree in Figure 5.) Communities for which the ratio of closed to open triads is too high are unlikely to grow. Although this shows up strongly in the data (see also Figure 6), it is not entirely clear how to interpret this result. It is possible that a large density of triangles indicates a kind of "cliqueishness" that makes the community less attractive to join; it is also possible that high triangle density is a sign of a community that stopped gaining new members at some point in the past and has subsequently been densifying, adding edges among its current set of members. We are currently pursuing further investigations to attempt to interpret the role of this feature more clearly.

4. MOVEMENT BETWEEN COMMUNITIES

Having analyzed the membership and growth of communities, we now turn to the question of how people and topics move between communities. A fundamental question here is the degree to which people bring topics with them from one community to another, versus the degree to which topics arise in a community and subsequently attract people from other communities. In other words, given a set of overlapping communities, do topics tend to follow people, or do people tend to follow topics? We also investigate a related question: when people move into a community are they more or less likely than other members of the community to be participants in current and future "hot topics" of discussion in that community?

While these questions are intuitively very natural, it is a challenge to define sufficiently precise versions of them that we can make quantitative observations. Furthermore, any attempt to make these questions precise will involve certain simplifications and approximations, and we start by discussing the reasons behind some of our experimental design decisions. We use the DBLP data discussed in earlier sections, with conferences serving as the communities (limiting the data to 87 conferences for which there is DBLP data over at least a 15-year time period). Since DBLP includes paper titles, we take the words in titles as the raw data for identifying topics in each community. There are a number of indications that the cumulative set of words in titles can serve, for our purposes here, as an effective proxy for top-level topics (see e.g. [22] and some of the discussion at the end of this section).

Informally, it is easy to think of individual instances where two conferences B and C seemed to move "closer together" over some period of years (for example, NIPS and ICML in the period 2000-2003 — an observation borne out by analysis of the data as well). We now define experiments that ask whether, in general over all such movement patterns, these movements are at the level of topics, people, or both — and if both, then which kind of movement tends to precede the other.

4.1 Time Series and Detected Bursts

Intuitively, it is possible for the same topic x to be "hot" at each of two conferences B and C at the same time, even if B and C are not highly similar in any "global" sense. Many of the effects we are seeking to understand have more the former flavor (a shared hot topic) than the latter (global similarity), so we structure our definitions around this former notion.

Term Bursts. For a given conference C and a word w, we denote by $T_{w,C}(y)$ the fraction of paper titles at conference C in year y that contain the word w. $T_{w,C}$ can thus be viewed as the time series giving the frequency of word w at C over a sequence of years. For each time series $T_{w,C}$, we identify bursts in the usage of w using a simple stochastic model for term generation that identifies intervals in which the usage can be approximated by a "burst rate" that is twice the average rate [22]. This burst detection technique was used in [22] on the same DBLP title data, and was observed to be effective at identifying "hot topics" at conferences. The same technique has since been used for finding term bursts in a range of other domains, for instance to detect current topics in blogs [24].

For our purposes, these burst intervals serve to identify the "hot topics" that indicate a focus of interest at a conference. We say that a word w is *hot* at a given conference C in a year y if the year y is contained in a burst interval of the time series $T_{w,C}$. (Note that being a hot term is a property of three things: a term, a conference, and a year.)

We also note an important caveat. Clearly it does not make sense to evaluate any *single* paper based on whether it happens to use a particular word in its title or not. All of our experimental findings based on burst analysis, however, only consider the frequencies of bursty words over large sets of papers, and will in all cases be supported by strong findings of statistical significance. In this way, the noise inherent in specific paper titles is being smoothed out by looking across large samples.

Movement Bursts. Next, we need to define a corresponding notion for author movement, and some care is needed here. Unlike title terms, individual people appear quite sparsely at conferences; even someone who is a "member" of a given conference community will generally not publish there every year. Moreover, movement is asymmetric — there may be movement from a conference B to a conference C but not vice versa — and so we need to employ a notion that is different from a simple overlap measure.

First, we define someone to be a *member* of a conference in a given year y if they have published there in the 5 years leading up to y. (In contrast to previous sections, this definition allows someone to be a member of a conference and later not a member, which is crucial for the kinds of analysis we do here.) We then say that author a *moves* into conference C from conference B in year y when a has a paper in conference C in year y and is a member of conference B in year $y-1$. Note that movement is a property of two conferences and a specific year, and further that although this measure of movement is asymmetric, it may sometimes hold in both directions.

Let $M_{B,C}(y)$ denote the fraction of authors at C in year y with the property that they are moving into C from B. Thus, $M_{B,C}$ can be viewed as a time series representing author movement, and we use burst detection to find intervals of y in which the value $M_{B,C}(y)$ exceeds the overall average by an absolute difference of .10.[5] We refer to such an interval as a $B \rightarrow C$ *movement burst*.

We now have word burst intervals, identifying hot terms, and movement burst intervals, identifying conference pairs B, C during which there was significant movement. We next discuss some experiments that investigate how these are aligned in time.

[5]We use an additive difference instead of a multiplicative factor to generate the burst rate here: multiplicative burst rates tend to penalize time series with large averages, and we need these here since they correspond to conference pairs with a large baseline overlap that nonetheless experience a sharp increase. While nearby values give similar results, we use a difference of .10 to define the burst rate since it produces about 200 burst intervals that are of moderate length, about 4 years each, over all conference pairs (B, C). For comparison, the word bursts average about 5 years in length.

	All Papers	Papers Contrib. to Movement
Num. papers	99774	10799
Currently hot	0.3859	0.4391
Future hot	0.1740	0.1153
Expired hot	0.2637	0.3102

Table 5: Fractions of papers containing hot terms. Papers contributing to a movement burst contain elevated frequencies of currently and expired hot terms, but lower frequencies of future hot terms.

4.2 Papers Contributing to Movement Bursts

We first consider characteristics of papers associated with some movement burst into a conference C; we find that they exhibit significantly different properties from arbitrary papers at C. In particular, one crucial difference is in the extent to which they use terms that are currently hot at C, and the extent to which they use terms that will be hot at C in the future. Given that movement bursts intuitively represent increased participation from some other community, these differences will provide a first perspective on the general question of whether topics are following people, or whether people are following topics.

We make this precise as follows. First, we say that a paper appearing at a conference C in a year y *contributes* to some movement burst at C if one of its authors is moving from some conference B into C in year y, and y is part of a $B \rightarrow C$ movement burst. These are precisely the papers that, intuitively, are part of the elevated movement from other conferences into C. Now, it is natural to ask whether these papers that contribute to movement bursts differ from arbitrary papers in the way they use hot terms. Here we say that a paper uses a hot term if one of the words in its title is hot for the conference and year in which it appears.

As a baseline, 38.59% of *all* papers use hot terms. (While this number is a useful benchmark for relative comparisons, its actual magnitude can clearly be affected by altering the settings of the burst detection parameters.) On the other hand, as shown in Table 5, 43.91% of all papers contributing to movement bursts use hot terms. This difference is statistically significant: if we consider a binary variable that is true $.3859$ of the time, then the probability of seeing a sample of size 10799 (the number of papers contributing to movement bursts) where the variable is true $.4391$ of the time is seen to be $< 10^{-15}$ using a Chernoff-Hoeffding bound.

Thus it is apparent that papers written by people who are part of a burst of authors moving into a conference are more likely to be about topics that are "hot", or experiencing a burst, than is the case for papers in general.

Given that papers contributing to a movement burst exhibit an elevated usage of hot terms, it is natural to also ask whether they also contain an unusually high proportion of terms that *will* be hot at some point in the future, or that *were* hot at some point in the past. Specifically, we say that a paper at a conference C in year y uses a future hot term if it contains a word that will experience a burst at C starting in some year $> y$; we say that it uses an expired hot term if it contains a word that experienced a burst at C ending in some year $< y$. As shown in Table 5, we find that papers contributing to movement bursts in fact use expired hot terms at a significantly higher rate than arbitrary papers at the same conference (31.02% vs. 26.37%), but use future hot terms at a significantly *lower* rate (11.53% vs. 17.40%). Again, these differences are statistically significant at comparable levels.

Taken together these results support the notion that a burst of authors moving into a conference C from some other conference B are drawn to topics that are currently hot at C; but there is also evidence that this burst of authors produces papers that are comparably impoverished in their usage of terms that will be hot in the future. In other words, any notion that they are "germinating" terms that will soon become hot at conference C is not borne out by the data; in fact, the opposite appears to be true.

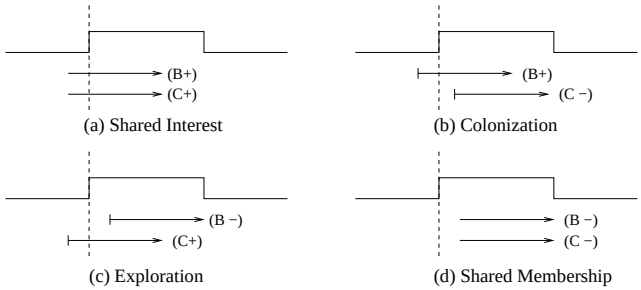

Figure 7: Four patterns of author movement and topical alignment: in each of (a)-(d), the labeled arrows represent term burst intervals for a shared hot term in conferences B and C, and the square wave represents a $B \rightarrow C$ movement burst. In the terminology from the text, (a) is shared interest, (b) is colonization, (c) is exploration, and (d) is shared membership.

We now turn to a second set of experiments that explores this temporal alignment of movement and term bursts in a different way, but leading to qualitatively similar conclusions.

4.3 Alignment between Different Conferences

We say that conferences B and C are *topically aligned* in a year y if some word w is hot at both B and C in year y. (We will also say that B and C are topically aligned *via* w.) Note that topical alignment, like movement, is a property of two conferences and a specific year. Also, two conferences can be topically aligned even if their overall collections of papers are quite different; they need only share a single common focus, in the form of a hot term.

It is natural to expect that two conferences are more likely to be topically aligned in a given year if there is also a movement burst going on between them. We first show that this is indeed the case, a basic result establishing that movements of terms and people are indeed correlated. Specifically, over all triples (B, C, y) such that there is a $B \rightarrow C$ movement burst containing year y, we find that 56.34% have the property that B and C are topically aligned in year y. As a baseline, only 16.10% of *all* triples (B, C, y) have the property that B and C are topically aligned in year y. Thus, the presence of a movement burst between two conferences enormously increases the chance that they share a hot term.

Given this, we are now in a position to ask one of the questions posed informally at the outset: do movement bursts or term bursts tend to come first? Specifically, whenever there is a $B \rightarrow C$ movement burst, we look at all hot terms w such that B and C are topically aligned via w in some year y inside the movement burst. There are now three events of interest:

(i) the start of the burst for w at conference B;
(ii) the start of the burst for w at conference C; and
(iii) the start of the $B \rightarrow C$ movement burst.

Let us consider how these might occur in order relative to one another, with interpretations of each; the various orders are depicted schematically in Figure 7. We then discuss how frequently these orders actually occur in the data.

- w bursts at both B and at C (in some order) before the $B \rightarrow$

	$C+$	$C-$
	(a)	(b)
$B+$	194 (0.6025)	32 (0.0994)
	(c)	(d)
$B-$	35 (0.1087)	61 (0.1894)

Table 6: Frequency of the four patterns relating movement and topical alignment. $B+$ (resp. $B-$) denotes that the burst of w at B precedes (resp. follows) the $B \to C$ movement burst; and analogously for C.

C movement burst begins. (See Figure 7(a).) We call this pattern *shared interest*, since the topical alignment of B and C happens before they come closer together in membership.

- w bursts at B, then the $B \to C$ movement burst begins, and then w bursts at C. (See Figure 7(b).) We call this pattern *colonization*, since one can imagine the movement from B to C as having a "colonizing" effect, carrying the term w from B (where it was already hot) to C (where it becomes hot).
- w bursts at C, then the $B \to C$ movement burst begins, and then w bursts at B. (See Figure 7(c).) We refer to this pattern as *exploration*, since one can imagine the hot topic at C attracting authors from B; subsequent to this "exploration" from B, the term becomes hot at B as well.
- The $B \to C$ movement burst begins, after which w bursts at B and at C (in some order). (See Figure 7(d).) We refer to this pattern as *shared membership*, since B and C come closer together in membership before the topical alignment happens via the common hot term w.

We now consider the relative frequencies of these four patterns. Over all cases in which there was a topical alignment of B and C concurrent with a $B \to C$ movement burst, we remove from the tabulation those in which two of the three relevant burst intervals (for the term at each conference, and for the movement) began in the same year. This leaves us with 322 instances in total, which are divided over the four categories as shown in Table 6. 194 of the instances correspond to the *shared interest* pattern: the term burst in each conference precedes the movement burst. In other words, of the four patterns, shared interest is 50% more frequent than the other three patterns combined. The next most frequent is shared membership, with 61 instances, followed by colonization and exploration with 35 and 32 respectively.

As with the previous set of experiments, we find that the intuitively appealing notion of authors from a conference B "transplanting" hot terms to a new conference C is not in fact the dominant type of movement in the data. Rather, it is much more frequent for conferences B and C to have a shared burst term that is already underway before the increase in author movement takes place.

5. CONCLUSIONS AND FURTHER DIRECTIONS

We have considered the ways in which communities in social networks grow over time — both at the level of individuals and their decisions to join communities, and at a more global level, in which a community can evolve in both membership and content. Even with very rich data, it is challenging to formulate the basic questions here, and we view the elaboration of further questions to be an interesting direction for future work.

The availability of complex datasets on communities in social networks, and their evolution over time, leads naturally to a search for more refined theoretical models. It will be interesting to connect standard theoretical models of diffusion in social networks to the kinds of data on community membership that one can measure in on-line systems such as LiveJournal. One class of questions was suggested at the end of Section 2 — forming accurate models for the asynchronous processes by which nodes become aware of their neighbors' behavior and subsequently act on it. Another goal is to understand how even very simple diffusion models may change if we parametrize influence not just by the number of neighbors who have adopted a behavior, but by the internal connectedness of these neighbors, following the findings in Section 2.

Finally, it would be interesting to relate some of the techniques developed here, particularly on movement between communities, to latent-space models for social networks as studied in Hoff et al. [18] and Sarkar and Moore [32]. Even without the network aspect, the movements in content exposed by very simple latent-space techniques are quite suggestive. For example, Figure 8 shows a representation of conferences from the DBLP dataset, encoded as term vectors and projected into a two-dimensional vector space X defined by Latent Semantic Indexing (LSI) [2, 10]. In each year, the set of conferences projects differently into X, and their collective motion over successive years provides some illustration of their changing relationships to one another. Such representations can clearly form the basis for alternate ways of quantifying community movement, with conferences forming natural groupings by topic, and with certain parts of the space becoming "filled out" as particular areas emerge over time.

Acknowledgements. We thank Ravi Kumar, Michael Macy, Eugene Medynskiy, and David Strang for valuable discussions, and Michael Ley for his generous help with the DBLP data.

6. REFERENCES

[1] Lada A. Adamic, Orkut Buyukkokten, and Eytan Adar, A Social Network Caught in the Web First Monday, 8(6), 2003.
[2] R. Baeza-Yates, B. Ribeiro. Modern Information Retrieval. Addison Wesley, 1998.
[3] S.Boorman, P.Levitt. The genetics of altruism. Acad. Pr. 1980.
[4] C. Borgs, J. Chayes, M. Mahdian and A. Saberi. Exploring the community structure of newsgroups Proc. 10th ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining, 2004.
[5] Leo Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.
[6] R. Burt. Structural Holes: The Social Structure of Competition. Harvard, 1992.
[7] D. Centola, M. Macy, V. Eguiluz. Cascade Dynamics of Multiplex Propagation. Physica A, to appear.
[8] J. Coleman Social Capital in the Creation of Human Capital, American Journal of Sociology, 94(Supplement):1988.
[9] J. Coleman. Foundations of Social Theory. Harvard, 1990.
[10] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman. Indexing by latent semantic analysis. J. Amer. Soc. for Information Science, 41(6), 1990.
[11] S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, A. Tomkins. Self-similarity in the Web. 27th International Conference on Very Large Data Bases, 2001.
[12] P.S. Dodds, D.J. Watts. Universal behavior in a generalized model of contagion. Phys. Rev. Lett., 92:218701, 2004.
[13] P. Domingos, M. Richardson. Mining the Network Value of Customers. Proc. 7th Intl. Conf. Knowledge Discovery and Data Mining,
[14] Gary Flake, Steve Lawrence, C. Lee Giles, Frans Coetzee. Self-Organization and Identification of Web Communities. IEEE Computer, 35:3, March 2002.
[15] G. W. Flake, R. E. Tarjan, and K. Tsioutsiouliklis. Graph Clustering and Minimum Cut Trees. Internet Math. 1(2004).

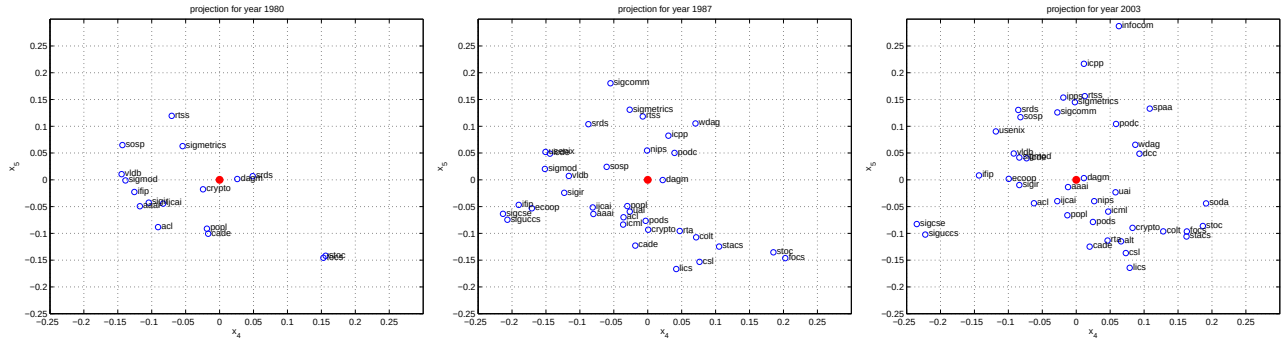

Figure 8: Projections of a subset of the conferences in three different years (1980, 1987, 2003) into a two-dimensional LSI space.

[16] M. Girvan, M. E. J. Newman. Community structure in social and biological networks. Proc. Natl. Acad. Sci. 99(2002).

[17] M. Granovetter. The strength of weak ties. American Journal of Sociology, 78(6):1360-1380, 1973.

[18] P. Hoff, A. Raftery, M. Handcock. Latent space approaches to social network analysis. Journal of the American Statistical Association, 97(2002).

[19] P. Holme, M. Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. arXiv physics/0603023, March 2006.

[20] J. Hopcroft, O. Khan, B. Kulis, B. Selman. Natural communities in large linked networks. Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining, 2003.

[21] D. Kempe, J. Kleinberg, E. Tardos. Maximizing the Spread of Influence through a Social Network. Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2003.

[22] J. Kleinberg. Bursty and Hierarchical Structure in Streams. Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2002.

[23] G. Kossinets, D. Watts. Empirical analysis of an evolving social network. Science, 311:88 – 90, 2006.

[24] R. Kumar, J. Novak, P. Raghavan, A. Tomkins. Structure and evolution of blogspace. Comm. ACM 47(2004)

[25] J. Leskovec, L. Adamic, B. Huberman. The Dynamics of Viral Marketing. Proc. 7th ACM Conf. on Electronic Commerce, 2006.

[26] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, A. Tomkins. Geographic routing in social networks. Proc. Natl. Acad. Sci. USA, 102(Aug 2005).

[27] M. Macy, personal communication, June 2006.

[28] M. E. J. Newman. Detecting community structure in networks. Eur. Phys. J. B 38, 321-330 (2004).

[29] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[30] M. Richardson, P. Domingos. Mining Knowledge-Sharing Sites for Viral Marketing. Proc. 8th Intl. Conf. on Knowledge Discovery and Data Mining. 2002.

[31] E. Rogers. *Diffusion of innovations* Free Press, 1995.

[32] P. Sarkar, A. Moore. Dynamic Social Network Analysis using Latent Space Models. SIGKDD Explorations: Special Edition on Link Mining, 2005.

[33] D. Strang, S.. Soule. Diffusion in Organizations and Social Movements: From Hybrid Corn to Poison Pills. Annual Review of Sociology 24(1998).

[34] T. Valente. *Network Models of the Diffusion of Innovations*. Hampton Press, 1995.

[35] F. Viegas and M. Smith. Newsgroup Crowds and AuthorLines. Hawaii Intl. Conf. Sys. Sci. 2004.

[36] X. Wang, A, McCallum. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends, Proc. 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2006.

[37] S. Wasserman and K. Faust. Social Network Analysis. Cambridge University Press, 1994.