Deconstructing centrality: thinking locally and ranking globally in networks

Sibel Adalı
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
Email: sibel@cs.rpi.edu

Xiaohui Lu
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
Email: lux3@cs.rpi.edu

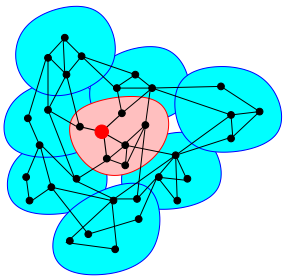
Malik Magdon-Ismail
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
Email: magdon@cs.rpi.edu

***Abstract*—We examine whether the prominence of individuals in different social networks is determined by their position in their local network or by how the community to which they belong relates to other communities. To this end, we introduce two new measures of centrality, both based on communities in the network: local and community centrality. Community centrality is a novel concept that we introduce to describe how central one's community is within the whole network. We introduce an algorithm to estimate the distance between communities and use it to find the centrality of communities. Using data from several social networks, we show that community centrality is able to capture the importance of communities in the whole network. We then conduct a detailed study of different social networks and determine how various global measures of prominence relate to structural centrality measures.Our measures deconstruct global centrality along local and community dimensions. In some cases, prominence is determined almost exclusively by local information, while in others a mix of local and community centrality matters. Our methodology is a step toward understanding of the processes that contribute to an actor's prominence in a network.**

Introduction

There are many algorithms for computing prominence, each operating on different sets of assumptions. For example, one family of algorithms [16] argue that it is not possible to measure an academician's prominence globally. According to these algorithms, prominence only makes sense in the context of a specific research community to which the researcher belongs. Alternatively, one can argue that researchers in core communities, i.e. those working on foundational problems are more prominent than the rest. How about researchers that serve as bridges between different communities, resulting in the transfer of ideas? Ultimately, these are all valid ways to define prominence. Different prominence measures, external or network based, incorporate these concerns to different degrees.

To effectively compute a *structural* prominence measure from the observed network interactions, we must understand the network factors that contribute to prominence. As a starting point, we have a network of actor-actor relations (for example, co-authorship on a publication, communicating with each other via blogs, friends on facebook, etc.). The basis for this research is that a typical social network contains social communities to which actors belong (an actor could belong to more than one community). A community is a subgroup of actors that are more closely related to each other than to the actors outside of the community. For simplicity, we assume that an actor belongs to just one community (this is a simplification in our analysis, but our methods readily generalize to when the communities are overlapping).



So, an actor (red node above) has a "status" within the communities to which it belongs, and the community itself has a "status" in relation to the other communities. The former we call the *local centrality*, and the latter the *community centrality*. The graph in Figure 1 below illustrates the notions.

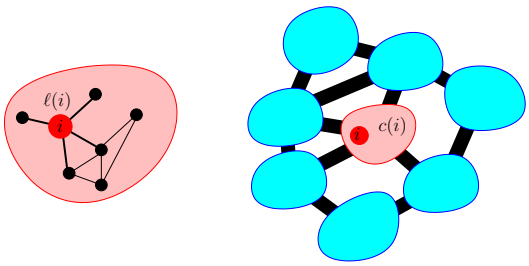


Fig. 1. (a) Node i has a *local centrality* $\ell(i)$ within its community. (b) Node i's community has a status within the "network" of communities, its *community centrality* $c(i)$.

Given a set of disjoint communities for a network, and an actor (node) i, we define two notions of centrality:

- **Local centrality** $\ell(i)$**,** which is a local measure of centrality with respect to only the nodes and links within the community. Any measure of centrality can be used to compute the local centrality, and for our study we tried closeness and betweenness [7], [17].
- **Community centrality** $c(i)$**,** which is a measure of centrality for node i's community. A community's

centrality (closeness or betweenness) is computed on a meta-network whose nodes are the communities, and the edges between communities indicate the 'distance' of the link between two communities. This meta-network needs to be computed from the underlying network and community structure, and we give one method to do so.

The community centrality captures global information regarding a node's community in relation to other communities in the whole network. Local centrality, on the other hand, considers centrality only with respect to one's local community. We can also define a global centrality for a node i, $g(i)$, for example the traditional closeness centrality, which uses structure in the entire network, ignoring community structure.

The goal of this research is to understand how these different measures of centrality contribute to the prominence of an actor. In particular, to show that each of the component parts into which we deconstruct centrality have *different* roles to play. Further, that these roles are different depending on:

- The measure of external prominence that one wants to capture. For example, with respect to bloggers, one can measure prominence as the sheer number of views a blogger receives; or the number of different (unique) users that the blogger attracts. The former captures the volume of interaction while the latter captures the size of audience.
- The role of the actor within a network. For example when an author in a collaboration network has low degree (versus high degree), then that author's local centrality may not be as important as its community centrality.

Our general approach is to use a linear model to explain prominence using various measures of centrality as the independent variables, for example:

$$\mathsf{prominence}(i) = w_\ell \cdot \ell(i) + w_c \cdot c(i) + w_g \cdot g(i) + \epsilon(i),$$

where $\epsilon(i)$ is an idiosyncratic noise. We use *cross validation* to study the significance of the regression coefficients (weights w_ℓ, w_c, w_g). That is, when does adding an independent regression variable help by lowering the out-of-sample prediction error as measured by leave-one-out cross validation. We use such a cross validation setting because it makes no distribution assumptions on the variables (such as Gaussian). We are indeed able to demonstrate, on a variety of social networks, that these different dimensions of centrality play very different roles.

Our Contributions

- Foremost, we introduce a new paradigm for measuring centrality that has two components: local and community. In order to compute these measures, we acquire a set of communities in the network. In this work we use the FastCommunity [4] community detection algorithm to obtain communities, but any method of choice for detecting communities is equally applicable. Indeed we illustrate the robustness of the results using a second community detection algorithm.
- Given communities (which we compute quickly using standard algorithms), our measures are more efficient to compute than global measures such as closeness which scale super-linearly. This is because we evaluate local centrality within a community, and communities are typically small; and, we evaluate community centrality using the community meta-graph, which is also typically a small graph. Hence our algorithms are nearly linear in the size of the network.
- We introduce a new algorithm to compute community centrality which uses the community structure to build a meta graph with communities as nodes and weighted edges between communities that capture the 'distance' between communities. We compute these weights between communities using a randomized algorithm.
- We study the role of our centrality measures in three real data sets: the DBLP academic publishing network; the network of actors in the movie and TV industry, IMDB; and, message data from an Irish forum. Our results demonstrate the expressive power of this new paradigm: different centrality measures are more important for different aspects of prominence, and for different types of nodes in the network. In some cases, they replace global centrality measures completely. There are many ways to implement our paradigm, in terms of how one computes local and community centrality but the message is that one's prominence is related in different ways to these different dimensions of structural centrality. In particular, local, community and global centrality measures are *all* different from one another.

Related Work

All commonly used measures of structural centrality are global in the sense that they use the entire network to capture how central a node is. Examples of such measures are closeness, degree, and betweeness centrality [3], [7], [17]. Some other measures that are based on random walks such as PageRank [10] or extensions of centrality based on all paths [15] and attention [1]. We use such a measure to compute local centrality but only using the subnetwork within a community. We also apply these measures to compute centrality of a community within the community meta-network.

It is widely accepted that communities exist [14], [17] and play an important role within social networks. However there is no systematic attempt to exploit this fact in computing measures of centrality. Two approaches to computing localized version of locality exist. In [16], the authors emphasize that comparing nodes in different academic communities is not very useful, and they show results on ranking nodes only within communities. In [11] global distances are computed up to a given bounded k. It has also been observed that global centrality alone does not capture important nodes in a socially driven networks, for example in airport networks [8] important airports may not be structurally central. There are also notions of centrality for a group [6] which define centrality

of a group with respect to the other *nodes* in the network. We have not found a notion of centrality for groups with respect to other groups, in particular illustrating how distances between clusters can be computed; we present one method for computing such measures of centrality based on a meta-graph of communities. As far as we know, there is no notion of community centrality comparable to ours, and there is no study that attempts to deconstruct prominence in terms of local and community centralities.

Community Based Centrality Measures

We consider networks of actors who are connected by virtue of interaction. For example, in the DBLP dataset, actors are authors of academic papers. There is a link between two authors, if they are co-authors on a paper. Similarly, in the IMDB dataset, actors are artists who star in movies and TV shows. Two actors are connected if they both starred in the same movie.

We represent the network as a simple graph $G = (V, E)$ where V is the non-empty set of nodes representing actors and $E \subseteq V \times V$ is the set of undirected edges representing interactions. The weights of edges represent the distance between a pair of actors. The more the actors interact with each other, the smaller is the distance. The distance $d(u, v)$ between two actors $u, v \in V$, is the length of a shortest path connecting the two nodes. We extend the notion of distance to a *restricted distance* $d_S(u, v)$ where $S \subseteq V$, which is the length of a shortest (u, v)-path that exclusively uses nodes in S. We extend the notion of distance to sets of nodes, $d(X, Y)$, where $X, Y \subseteq V$ are sets of nodes. The distance $d(X, Y)$ is the average of $d_{X \cup Y}(x, y)$ over pairs of nodes $x \in X, y \in Y$.

$$d(X, Y) = \frac{1}{|X| \cdot |Y|} \sum_{x \in X, y \in Y} d_{X \cup Y}(x, y)$$

Community Graph and Centrality

Let $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ be a set of communities, where each $\mathcal{C}_i \subseteq V$ is a community (group of nodes). For simplicity we assume that $\mathcal{C}$ is a partition of the nodes (a disjoint cover), so the communities are non-overlapping. In all the networks we study, we use the FastCommunity [4] community detection algorithm based on the modularity principle for discovering the communities. However, we have also run comparisons with a different community detection algorithm [12] to test the robustness of results.

Given a set of communities, we define a **community meta graph** $\widetilde{G}(\mathcal{C}) = (\widetilde{V}, \widetilde{E})$ where the nodes represent communities and the edges represent the connectivity between communities. The graph is constructed as follows:

1) For each community $\mathcal{C}_i \in \mathcal{C}$, we create a node $\widetilde{v_i} \in \widetilde{V}$.
2) An edge $(\widetilde{v_i}, \widetilde{v_j})$ is created if there are two nodes $x, y \in V$ such that $x \in \mathcal{C}_i$ and $y \in \mathcal{C}_j$ and $(x, y) \in E$.
3) Edge weights (distances) between communities are determined by computing the average restricted distance between nodes from the two communities. Specifically, given $(\widetilde{v_i}, \widetilde{v_j})$, the edge weight between the corresponding communities is:

$$w(\widetilde{v_i}, \widetilde{v_j}) = d_{\mathcal{C}_i \cup \mathcal{C}_j}(\mathcal{C}_i, \mathcal{C}_j)$$

Intuitively, the community graph is a meta-graph with communities as nodes and all edges between two communities being condensed into a single weighted edge. The edge weight between the communities depends on the distance between all pairs of nodes from the two communities where distance is measured in the subgraph induced by the communities.

As computing the average distance between two communities can be quadratic, we use a random sampling algorithm to estimate it. Let $\mathcal{S}_i \subset \mathcal{C}_i$ be a random sample of nodes from $\mathcal{C}_i$, where a node is sampled with probability p_i. Let $\alpha_i = |S_i|/|C_i| \approx p_i$ be the fraction of nodes sampled. We use a sampling based estimate of $w(\widetilde{v_i}, \widetilde{v_j})$ given by:

$$\widehat{w}(\widetilde{v_i}, \widetilde{v_j}) = \frac{\alpha_i \cdot d_{\mathcal{C}_i \cup \mathcal{C}_j}(\mathcal{S}_i, \mathcal{C}_j) + \alpha_j \cdot d_{\mathcal{C}_i \cup \mathcal{C}_j}(\mathcal{C}_i, \mathcal{S}_j)}{\alpha_i + \alpha_j}$$

For smaller communities, we use a larger sampling probability to preserve accuracy. The role of sampling is to simply improve efficiency. We have found that the specific form of distance measurement does not play a large role, so an approximation suffices. Aside from average distance as a measure be community-community weight, we have also tried minimum and maximum distance and our results are robust to such choices, so we do not report on them.

Given a community meta-graph as computed above, we may now compute centrality measures on this meta-graph, which in turn give the community centralities of the nodes within the communities.

Experimental Setup

We study a number of centrality measures using different data sets. The global centrality measures are used for comparison with the newly introduced local and community centrality measures. We summarize our centrality measures below, and Table I is a useful reference for the notation.

- **Degree Centrality.** A node's degree in G, normalized by $|V - 1|$, denoted by **deg**.
- **Global Centrality.** Global closeness centrality (**cc**) is the inverse of a node's average distances to all the other nodes. Global betweenness centrality (**bc**) is the average of fractions of a node lie on a shortest path between all the possible pairs of nodes.
- **Local Centrality.** Local closeness (**lcc**) and local betweenness (**lbc**) are the closeness and betweenness centrality on the subgraph induced by a community of nodes respectively.
- **Community Centrality.** Community closeness (**ccc** and community betweenness (**cbc**) centrality are the closeness and betweenness centrality on the meta-network of clusters respectively.

Datasets

DBLP (Digital Bibliography & Library Project): is a dataset containing information about scientists (actors) from

Measure	Name
deg	Degree centrality
cc	Closeness centrality
bc	Betweenness centrality
lcc	Local closeness centrality
ccc	Community closeness centrality
lbc	Local betweenness centrality
cbc	Community betweenness centrality
size	Size of actor's community

(a)

Measure	Name	Dataset
h	H-Index	DBLB
t	TC-10	DBLP
budget	average movie budget for actors	IMDB
gross	average movie gross for actors	IMDB
rating	average movie rating for actors	IMDB
view	total views for a thread	boards.ie
audience	number of distinct posters for a thread	boards.ie

(b)

TABLE I. THE LIST OF (A) CENTRALITY AND (B) GROUND TRUTH MEASURES STUDIED IN OUR PAPER.

Computer Science, their publications (objects) and the publication venues[1]. Our data set consists of 615,416 authors (actors) and 2,323,509 edges.

IMDB (Internet Movie Database): contains information about the movie industry in general[2]. Note that IMDB contains information from multiple movie industries. We limit ourselves to only movies made in the USA. From this data, we extract information about movie stars (actors) who star in movies (objects) as well as directors who direct movies. We examine the IMDB data in decades as some of the prominence measures we study in the next sections are only meaningful in a small window of time. Note that, for each movie we choose the top three actors based on the billing order to separate actors with significant roles from the others.

Decade	Actors	Edges	Movies	Budget Info	Gross Info	Rated movies
1930s	5723	40145	10285	411	72	5789
1960s	4831	17039	3787	348	122	3325
2000s	32557	82832	18633	8089	3080	13059

TABLE II. NUMBER OF US MOVIES WITH BUDGET, GROSS AND RATING INFORMATION IN IMDB AND THE SIZE OF THE GRAPH FOR EACH DECADE.

boards.ie (Irish Forum Dataset): contains ten years of forum discussions from 1998 to 2008, containing around 9 million documents[3]. It contains posts organized into threads of discussion, authors and FOAF data for the users. We consider a reply to a post as an interaction between the creator of the post and the author of the post that is being replied to. To reduce the size of the graph, we remove all actors with only one post and also actors with more than 3 standard deviations of posts (1850+) as such actors tend to be moderators. Based on this, we construct a graph of posters, containing 64,579 actors and 2,153,832 edges.

In these datasets, we construct an actor-actor graph, in which the nodes are people. Two people are connected if they have collaborated on a paper, a movie or a thread. The weight of an edge (u, v) is determined by:

$$w_a(u,v) = \frac{1}{\sum_{o \in \Gamma(u) \cap \Gamma(v)} \frac{1}{log(|\Gamma(o)|)}}$$

where $\Gamma(u)$ is the set of objects that actor u has created, and $\Gamma(o)$ is the set of actors who have collaborated on object o. This variation of the Adamic/Adar [2] measure looks at the common objects between actors, such as the common papers. The more such objects there are, the smaller is the distance. However, a collaboration on an object is more valuable, if there are not many other collaborators on it, given by $\Gamma(o)$. This measure of attention becomes more important for DBLP. In IMDB, we fix the number of actors for each movie to be around three.

Given an actor- actor network, we compute communities using the FastCommunity [4] community detection algorithm based on modularity, then compute community distances and the community centrality.

Ground Truth Values Used in Our Tests

To understand the value of our new measures of centrality, we relate them to external non-structural measures of prominence (ground truth).

DBLP: In DBLP, a researcher's prominence is based on the amount of citation her papers get. We consider two different measures based on citations:

- **h:** The H-Index [9] of an author is h if h of her papers received at least h citations each, and each of the rest has at most h citations.
- **t**: The TC-10 value of an author is the average number of citations of the author's top 10 most cited papers.

IMDB: In IMDB, the prominence of actors is generally tied to the success of their movies. There is not a single measure of success. We look at multiple for an actor: average movie budget **(budget)**, average movie gross **(gross)** and average movie rating **(rating)** in a specific decade.

Movie gross is arguably a noisy measure of prominence as it is notoriously hard to predict which movies will gross well [5]. Furthermore, many other factors such as marketing and herd behavior [13] play a significant role in a movie's success at the box office. Movie budget is a measure of how much the movie industry believes that an actor will produce a successful movie. Movie budgets are also noisy as a significant portion may be allocated to other factors such as special effects and marketing in some movies, and to actors in others. The third measure the overall rating of the movie, while subjective, shows the value of the movie in terms of the audience satisfaction. For this, we use the rating information in IMDB.

For budget and gross values, we introduce a normalized measure to reduce the noise in the actual values. We first partition the movies into decades. We then rank movies by their budget (or gross) within the decade it belongs to. We assign a value to the movie i (normalized movie value) by the equation:

$$mv(i) = \frac{k - r(i)}{k - 1}$$

[1] www.informatik.uni-trier.de/∼ley/db/

[2] www.imdb.com

[3] http://www.icwsm.org/2012

Fig. 2. Common words in highly central communities in DBLP.

where $r(i)$ is the rank of movie i, and k is the total number of movies in that decade. The prominence of an actor in a specific decade is given by the average value of her movies given by the specific measure. For each decade, we only consider the actors who were active in that decade and compute centrality values for the movie graph of that decade. The rating information is not normalized, it is a value between 1 and 10.

boards.ie: We consider the total number of views a thread has accumulated (**view**) and the total number of distinct people who have participated in a given thread (**audience**) as the ground truth for a thread. For each person, we average the two statistics for the threads they have originated.

Understanding Community Centrality

DBLP: We first study the meaning of community centrality. To this end, we first look at the communities for DBLP. We look at the largest communities at the two ends of the spectrum, highest ranked communities (sizes around 15K actors) and lowest ranked communities (sizes around 1K actors). We look at the venues (conferences and journals) for all the publications of all the actors in a community. We treat the words for each actor as a document and extract the terms from these after removing any stop words. We then adjust the frequency of each word with the usual TF-IDF [18] measures within the community (which devalues very common words like conference and international). Using these weighted frequencies we construct a word cloud.

The results for closeness centrality are given in Figure 3. One thing we notice is that central communities have terms that correspond to very high level terms like Artificial Intelligence, Databases, Programming and Multimedia. One can consider these communities as containing researchers doing the most mainstream and foundational research. One can expect that the research in these areas impact research in many more applied research areas. More peripheral communities on the other hand use more specialized terms such as microelectronics, bioinformatics, circuits, wireless and neural. One can visualize that words in the central communities correspond to concepts that are more general than those in less central communities.

IMDB: Unfortunately, no similar concept of venues or general concepts exist for the movies to understand the communities in IMDB. Instead, we consider the popularity of actors in general which we find by querying the actor's full name in Google[4]. A popular actor is likely to have a lot more hits for their name than a less popular actor. To do so, we pair actors from two different communities: actor A_i from community $\mathcal{C}_i$ and actor A_j from community $\mathcal{C}_j$ such that actors A_i and A_j have similar numbers of movies, communities $\mathcal{C}_i$ and $\mathcal{C}_j$ have similar sizes, but $\mathcal{C}_i$ is much more central than $\mathcal{C}_j$ (the rank of the two is separated by at least 100 communities among the 368 in our results). We also consider the large communities in our data set. From the set of all possible pairs, we sample about 10% randomly.

We then find the number of query results for each term A given by freq(A), and compute freq(A_i)/freq(A_j) for all the pairs we study. The results are in the range between 0.0004 and 80,568 with average 237 and median 1.2. So, on the average, an actor from a more central community is 237 times more popular than an actor from a less central community. It seems there is some support that actors from more central communities are more mainstream compared to those in the less central communities. However, given the median is 1.2, the picture is more complex indicating that the average may be getting skewed by extremely popular actors.

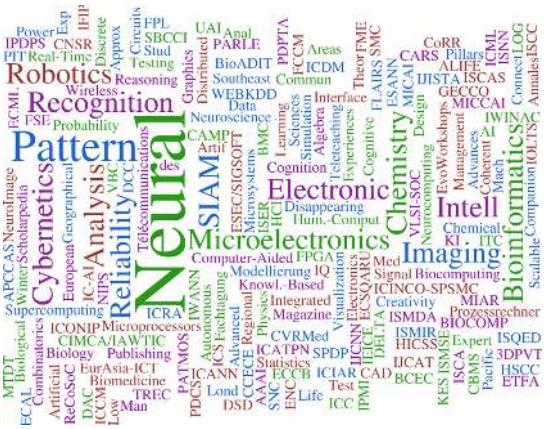
boards.ie: Given that central communities are those that represent the most general interests in that network, we apply the same process to the top 10 communities in the boards.ie dataset according closeness centrality. The top terms in this network are shown in the table below. The top interests are mostly related to computing and to some degree gaming. This correlates well with the main audience of this network as it is described on other sites on the Internet.

Rank	Terms
1	laptop pc game time wireless player sky music xbox
2	car pc broadband laptop nokia wireless dvd eircom tv phone
3	broadband pc eircom game tv dvd laptop wireless nokia player
4	noah sylvan matter warning jungle debate
5	broadband pc game player xbox airsoft laptop wireless tv
6	asia summer recognise student australia table japan meeting tennis travel
7	poker hand game online play tournament car card broadband boards
8	cork thread tralee car driving city bang broadband road
9	skateboard aerial 802.11g food veggie juggling avi alternative pcmcia
10	balbriggan northern goss major scam end house hard moved private

TABLE III. Top terms used in the most central communities in the boards.ie dataset.

[4]http://www.google.com

Fig. 3. Common words in peripheral communities in DBLP.



COMPARISON OF PROMINENCE MEASURES

We now study the impact of local and community centrality on prominence in general. We divide our features into two sets: local and global. Global features (G) are the well-known global centrality measures: **deg**, **cc**, **bc**. The local features (L) are given by **ccc**, **lcc**, **cbc**, **lbc**, **size**. Note that we have added size as we have abstracted it out in the normalization process. We consider two separate questions:

- L→G: If we are given the local features, do the global features improve the prediction accuracy further?
- G→L: If we are given the global features, do the local features improve the prediction accuracy further?

To compute this, we use a two-step forward subset selection based regression (FSS) using cross validation error as our criterion for adding a feature in the step regression. In each step, we find which of the input features improve the prediction accuracy in a linear step-wise fashion. To account for bias, we add a constant factor, **1** to all runs. For L→G, we first find which of the local features are best predictors. Then, we run FSS again with both L and G features. This time, we require FSS to use the features found in the previous run. This computation finds which global features improve on the existing local features. We then select all the features that pass our significance criteria and report on those. Even though some features were used, they may not appear in the results if they do not pass the significance criteria. The reverse is performed for G→L, first finding features for G and then requiring them to exist in the second run including all the features.

The FSS method performs regression on an input matrix X, in our case all the centrality values, and a target vector y, in our case a ground truth value for each actor. The result of FSS is a weight vector w that best predicts y with $X^T w$. However instead of computing a weight for each feature which may result in overfitting, we use a greedy forward stepwise regression to minimize the leave-one-out cross validation (LOO-CV) error. At each step, the process builds on already selected features from X. When choosing the $(k+1)^{th}$ feature, the LOO-CV error is computed assuming the previous k features are already selected. If the LOO-CV prediction error decreases with the $k+1^{th}$ feature, then the features is added. Otherwise the process stops and we output the sparse regression vector w using only the k selected features. Note that we normalize all features separately to make it possible to compare weights across different experiments.

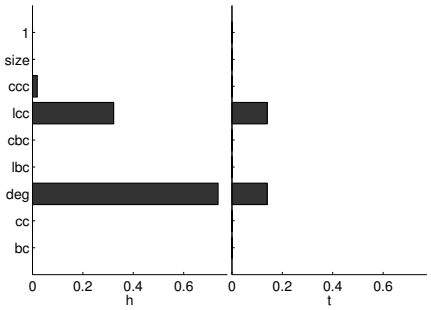


Fig. 4. Weights for predicting H-Index (h) and TC-10 (t) in DBLP data. Light bars indicate negative weights and dark bars indicate positive weights (L→G).

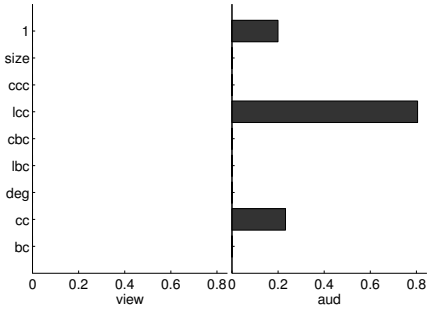


Fig. 5. Weights for predicting views and audience in boards.ie data. Light bars indicate negative weights and dark bars indicate positive weights (L→G).

DBLP: The most predictive features are shown in Table IV and the weights are shown in Figure 4. We see that degree is by far the most predictive feature in this dataset. The more actors that you are connected to, the better social capital you have. This is true because you get more information from the network and at the same time more people know and cite your work. For H-Index, community closeness is more important as work in more foundational areas tend to get cited more widely leading to higher H-Index values. However, for outlier behavior measured in TC-10, local factors like the community size and

H-Index	
L→G	G→L
ccc^{***}	ccc^{***}
lcc^{*}	deg^{***}
deg^{***}	

(a) DBLP

TC-10	
L→G	G→L
lcc^{***}	$size^{***}$
deg^{*}	deg^{**}

budget	
L→G	G→L
1^{***}	1^{***}
ccc^{***}	ccc^{***}
cc^{***}	cc^{***}

gross	
L→G	G→L
1^{***}	ccc^{***}
cbc^{***}	lcc^{***}
cc^{**}	cc^{***}

(b) IMDB (2000s)

rating	
L→G	G→L
1^{***}	1^{***}
$size^{***}$	$size^{**}$

audience	
L→G	G→L
1^{***}	1^{***}
lcc^{***}	cc^{***}
cc^{*}	

(c) boards.ie

TABLE IV. THE MOST PREDICTIVE CENTRALITY FEATURES FOR ALL THE DATASETS, PRESENTED IN THE ORDER OF IMPORTANCE. **1** REPRESENTS THE CONSTANT FACTOR. COMMUNITIES ARE DETECTED BY THE FASTCOMMUNITY [4] ALGORITHM. DISTANCE BETWEEN TWO COMMUNITIES ARE COMPUTED BY AVERAGING DISTANCES OF RANDOM SET OF NODES IN THE TWO COMMUNITIES. FOR EACH FACTOR, WE USE * FOR SIGNIFICANCE AT 10%, ** FOR SIGNIFICANCE AT 5%, AND *** FOR SIGNIFICANCE AT 1%. NOTE: NO FACTOR IS FOUND TO BE SIGNIFICANT FOR THE VIEW MEASURE IN BOARDS.IE.

the local centrality play a bigger role. If you have ground breaking work, the people in your community will appreciate it regardless of where the community lies.

IMDB: In Table IV and Figure 6, we track the change in the prominent features across different decades. Given that we have more data in later decades, the results are more likely to be representative of the movie industry in these decades. In IMDB, the constant factor is quite significant. Hence, all predictions include a prediction based on the average actor in the database. In particular, the ratings are highly biased toward generally positive due to their self-selective nature: people will rate a movie if they like it. As a result, we only found community size significant for this measure on top of the constant factor. In fact, IMDB contains one very large cluster that contributes to this result.

For budget, clearly both global and community closeness are very significant. This means that one's standing in the network as a whole and the importance of community together are very important. Clearly, one's place in the network plays an important role in getting chosen to be a part of high budget movies. This holds true for most of the later decades in IMDB. For gross, the picture is less clear. Being in high betweenness communities (*cbc*) is a factor, which could mean that actors in this group are known to a larger group of people due to their versatility. In fact, *cbc* is a factor also in ratings in previous decades. In later decades, global closeness centrality (*cc*) becomes more important for gross. One explanation could be that the highest grossing films and the highest budget films are more and more correlated as studios invest heavily in some movies. As a result, global closeness centrality is a factor in both.

boards.ie: Finally, In Table IV and Figure 5, we analyze the boards.ie dataset. One expects that this is one of the most noisy data sets. As a result, there are no factors for the number of views. For audience, local closeness centrality (*lcc*) and closeness centrality (*cc*) are both positive indicators and can be substituted for each other. However, *cc* is slightly more important as there is a value in having a global presence in the network. However, there is no coherent community influence in this dataset.

Robustness: To summarize, we see that *cc*, *lcc* and *ccc* are all distinct factors, providing different network level information. We have shown that some networks have strong community based prominence measures that are better captured with the existence of our community based factors. In fact, these factors significantly improve prediction over the global factors. This is not true in all networks however, as we have shown in boards.ie. Our local and community based measures are cheaper to compute than the global measures as they work on reduced networks and provide novel ways to measure prominence.

We end this section with a discussion of the robustness of the results for different community detection algorithms. We have applied the Walktrap algorithm [12] to all our datasets which is based on a concept of a random walker who gets trapped in dense areas of the network. We are showing results for IMDB only for space reasons in Figure V. Almost all the results remain the same, but there is a small difference regarding the community size feature. Density is more of a global feature as opposed to the modularity computed by FastCommunity which is a local feature. As a result, community size is not a significant factor in ratings and contributes slightly to budget. Overall, the results do not change much due to the choice of the community detection algorithm. However, choosing an algorithm that is based on a local criteria is more desirable in general as local centrality in this case is more meaningful.

budget	
L→G	G→L
1^{***}	1^{***}
ccc^{***}	ccc^{***}
$size^{***}$	cc^{***}
cbc^{*}	
cc^{***}	

gross	
L→G	G→L
1^{***}	ccc^{***}
ccc^{***}	lcc^{***}
cc^{**}	cc^{***}

rating	
L→G	G→L
1^{***}	1^{***}

TABLE V. THE MOST PREDICTIVE CENTRALITY FEATURES FOR THE IMDB 2000S DATASET WITH COMMUNITIES DETECTED BY THE WALKTRAP [12] ALGORITHM.

CONCLUSIONS

In this paper, we presented a new way to look at centrality. Instead of considering the centrality of actors in the whole network, we look at their centrality within their own community and the centrality of their community within the whole network of communities. We investigated when local and community centrality measures matter, and whether these deconstructed centrality measures replace the well-known centrality measures. To test the efficacy of our measures, we studied three large networks: academic paper publishing, movie industry and an Irish message board.

Our findings suggest that our measures are significant indicators of many different measures of prominence. In many cases, they complement and significantly improve on the global centrality measures. However, their importance vary depending on the ground truth measures and networks considered. There needs to be an underlying community structure for these measures to be important. In measures like H-Index for academic publishing and movie ratings, there is a certain expectation that prominent actors must come from a community that is fundamental in some way. In the academic network, we have seen that central communities revolve around topics that are foundational and any one in the network is likely to be familiar

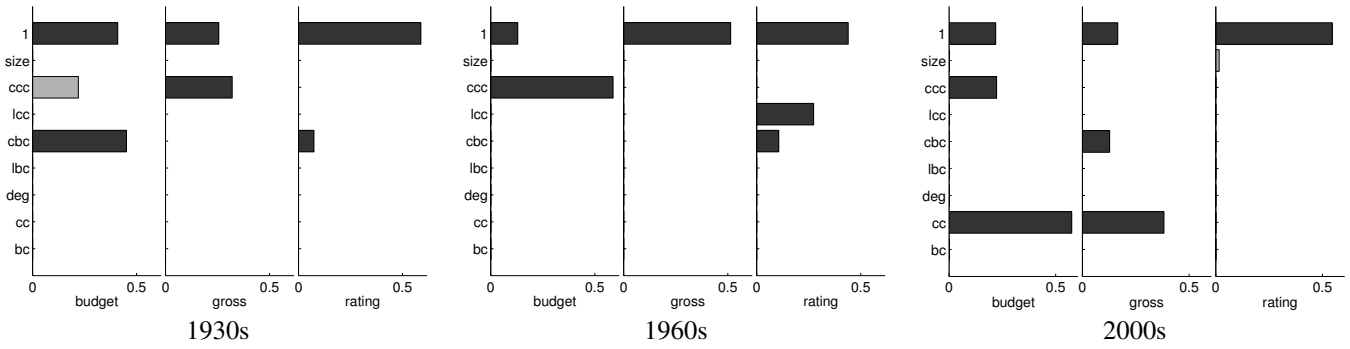


Fig. 6. Weights for predicting budget, gross and rating in IMDB data from 1930s, 1960s, 2000s. Light bars indicate negative weights and dark bars indicate positive weights (L→G).

with these topics. In the movie industry, central communities contain actors who star in high budget movies which cater to the tastes of the mainstream audience. As a result, community based centrality measures emerge as strong indicators for associated prominence measures.

Our deconstruction also allows us to study whether actors' prominence within their community or in the whole network play an important role in their prominence. Local importance suggests that prominence measures are based on local processes. Global closeness plays a role in cases where an actor needs to be a superstar in the whole network to be prominent. For example, in the movie database, this is true especially in later decades. Star driven blockbuster movies with an expected audience are used more and more frequently as a way to manage the inherent uncertainty of the film industry [5]. Our measures provide a way to better tune the structural analysis of networks at various levels of granularity.

Many interesting problems remain. We would like to study the predictive power of different measures for different partitions of data (high vs. low degree actors and small vs. large communities). We also would like to study the characteristics of central and peripheral communities. We would like to analyze communities in the Internet movie database through other means, and also investigate the distinction between community based closeness and betweenness measures. The betweenness measures suffer in smaller networks as there are many nodes that do not lie on any shortest paths and have betweenness of zero. More robust versions of betweenness can be used here to better understand the impact of community level betweenness for determining prominence. We hope to apply this type of analysis to many other networks and gain further insight into prominence in these networks.

ACKNOWLEDGMENTS

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] S. Adalı, X. Lu, and M. Magdon-Ismail. Attentive betweenness centrality (abc): Considering options and bandwidth when measuring criticality. In *2012 ASE/IEEE International Conference on Social Computing*, 2012.

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

[3] S. P. Borgatti and M. G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466 – 484, 2006.

[4] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys Rev E*, 70(6):066111+, 2004.

[5] A. De Vany. *Hollywood economics: How extreme uncertainty shapes the film industry*. London: Routledge, 2004.

[6] M. Everett and S. P. Borgatti. *Extending Centrality*. Cambridge University Press, 2005.

[7] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[8] R. Guimera and L. Amaral. Modeling the world-wide airport network. *Eur. Phys. J. B*, 38:381–385, MAR 2004.

[9] J. Hirsch. An index to quantify an individual's scientific research output. *Proc. of the National Academy of Sciences*, 46:16569–16572, 2005.

[10] A. Langville and C. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1:335–380, 2005.

[11] J. Pfeffer and K. Carley. k-centralities: Local approximations of global measures based on shortest paths. In *Proceedings of WWW 2012 LSNA'12 Workshop*, pages 1044–1050, 2012.

[12] P. Pons and M. Latapy. Computing communities in large networks using random walks. In *Proc. 20th Comp. and Inf. Sc.*, pages 284–293, 2005.

[13] M. J. Salganik and D. J. Watts. Web-based experiments for the study of collective social dynamics in cultural markets. *Topics in Cognitive Science*, 1(3):439–468, 2009.

[14] J. Scott. *Social network analysis*. SAGE Publications Limited, 2012.

[15] K. Stephenson and M. Zelen. Rethinking centrality: Methods and examples. *Social Networks*, 11(1):1–37, Mar. 1989.

[16] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proc. 15th SIGKDD*, pages 797–806, 2009.

[17] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[18] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3):13:1–13:37, June 2008.