Social Network Analysis and Mining manuscript No.
(will be inserted by the editor)

Local, community and global centrality methods for analyzing networks

Sibel Adalı · Xiaohui Lu · Malik Magdon-Ismail

Received: date / Accepted: date

Abstract We examine whether the prominence of individuals in different social networks is determined by their position in their community, the whole network or by the location of their community within the network. 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 central communities incorporate actors who are involved in mainstream activities for that network. We then conduct a detailed study of different social networks and determine how various global measures of prominence relate to structural centrality measures. We show that depending on the underlying measure of prominence, different combinations of local, global and community centrality play an important role in determining the prominence. Local and community centrality measures add new information on top of existing global measures. We show robustness of our results by studying different partitions of the data and different clustering methods. Our deconstruction of centrality allows us to study the underlying processes that contribute to prominence in more detail and develop more detailed and accurate models.

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

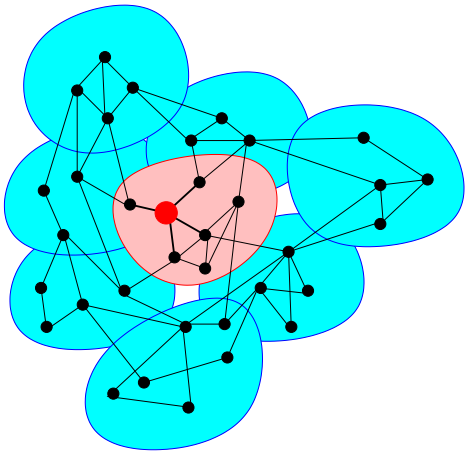
Xiaohui Lu
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
lux3@rpi.edu

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

Introduction

There are many algorithms for computing prominence, each operating on different sets of assumptions. For example, one family of algorithms [22] 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. External or networked based prominence measures are typically based on different social processes that contribute to prominence. To effectively compute a *structural* prominence measure from the observed network interactions, we must understand these processes clearly. That is what we aim to enable with the new centrality measures that we introduce in this paper.

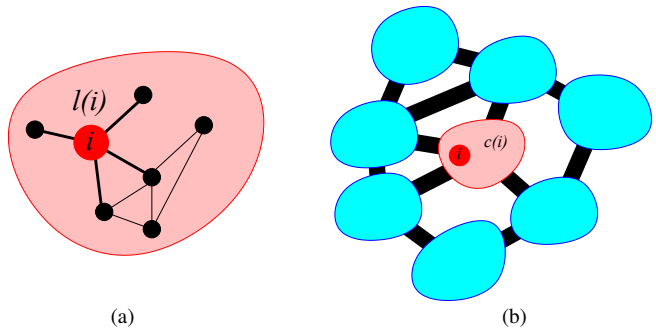
As a starting point, we have a network of actor-actor relations (for example, co-authorship on a publication, communicating with each other via blogs, friends on facebook, etc.). The basis for this research is that a typical social network contains social communities to which actors belong (an actor 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). An example community structure is shown below.



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*. For example, an actor could be a member of a prominent community but be a small player in that community. In this case the actors community centrality would be high but local centrality low. The graph in Figure 1 illustrates the notions.

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 [9, 24].



(a) (b)

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* $c(i)$.

– **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 her 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 [5] 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. We study various prominence measures for each data set. 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.
- We illustrate the robustness of our results by studying different partitions of the data with different characteristics. We investigate different linear models by requiring different centrality measures to be present in the model, and evaluate the fit of different linear models using Kendall-tau rank correlation. We show that in many partitions, local and community based centrality measures remain crucial in predicting prominence. These results replicate those found for the whole network.

Through our study, we gain insight into how the same centrality measure can take a different meaning at different levels. Global closeness centrality is especially useful for finding the stars in the whole network: individuals who are well-known to almost everyone. Local closeness centrality captures individuals who are embedded in a specific community. For example, to have a highly cited paper, an academician has to be known within her own discipline. Community closeness centrality reveals the foundational or mainstream activities in the network. Individuals from centrality communities tend to make impact in many communities. To have many papers with high citations, it is important to be in a foundational field from which ideas can diffuse to other more applied fields.

The implementation of the community centrality algorithm used in this paper is available as open source code at `github.com/rpitrust/prominence`.

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 [4, 9, 24]. Some other measures that are based on random walks such as PageRank [14] or extensions of centrality based on all paths [21] and attention [1]. These measures all use some form of criticality of a node to the paths in a network. Since all paths between all pairs of nodes may be considered, these measures are global. We use such a measure to compute our local centrality measure, but the crucial point is that we use only the *local* subnetwork relevant to the node that is defined by the node's community. We also apply these measures to compute centrality of a community within the community meta-network.

It is widely accepted that communities exist [20, 24] 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 centrality exist. In [22], 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 [16] global distances are computed up to a given bounded distance k. It has also been observed that global centrality alone does not capture important nodes in socially driven networks, for example in airport networks [11] important airports may not be structurally central. There are also notions of centrality for a group [8] which defines centrality of a group with respect to the other *nodes* in the network. This notion of "group-centrality" directly extend a nodes centrality by positing that a group is central if any one of the nodes in the group are central. For example the extension of node-betweeness-centrality to group-betweenness-centrality is to compute the number of shortest paths that us a node in the group. A similar approach has been proposed for biological networks [26], where the nodes that lie in central modules are shown to perform basic metabolic functions, while peripheral nodes provide more specialized functions. 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.

We also provide a novel study that shows the differences between local, global and community based centrality measures. We show the characteristics of central communities and illustrate that the three different notions of centrality are complementary using multiple datasets and external measures of prominence. In this paper, we extend our earlier work [2] and conduct a thorough study of the robustness of our findings. We show that deconstruction of centrality measures is useful not only for the whole network, but also for more homogeneous subsets of the data. Even when we consider a subset of the individuals in the network, such as those with high (or low) degree, with many (or few) artifacts, the local and community based centrality measures remain crucial for differentiating between the prominence of these individuals.

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 [5] community detection algorithm based on the modularity principle for discovering the communities. However, we have also run comparisons with a different community detection algorithm [17] 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:

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

   (Implicit inour definition is that nodes in a community are connected, that is the subgraph induced by the vertices in a community is connected.)

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 1 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.

Clearly, these are not the only centrality measures available. Our aim is not to study all possible centrality measures, but a way to decompose the existing measures. As such, we chose the three most commonly used measures; degree, closeness and betweenness. Pagerank or eigenvector centrality is yet another frequently used measure. In the datasets we analyzed here, pagerank and degree were highly correlated and provided very similer results. As a result, we have decided to include only one of the two.

Note that to compute the closeness centrality for all the nodes, it is equivalent to solving the all pair shortest paths problem (APSP). The APSP problem can be solved by various algorithms in either $\mathcal{O}(|V||E| + |V|^2 log|V|)$ using the Dijkstra's algorithm for weighted graphs [6] or $\mathcal{O}(|V||E|)$ using the Floyd-Warshall algorithm for unweighted graphs [6]. Betweenness centrality has the same overall complexity. As a result, the running time is very closely related to the graph size. To give an example, we ran experiments on a Dell workstation with i7 2.8GHz CPU (4 physical cores), and 12 GB memory. The OS was Ubuntu 10.04,

and we also used the boost library (http://www.boost.org/) for network manipulation. For the DBLP dataset described below (which is our largest data set), clustering step took 6-8 hours. Computing community distance took 20 - 24 hours, which is dependent on the connectivity of communities. DBLP had the highest number and the most densely connected communities. Global betweenness and closeness also took many days to run in contrast with local centrality that took around 1-2 hours. Overall, local and community centrality computations both provided significant speed up over the global methods.

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 1 The list of (a) centrality and (b) ground truth measures studied in our paper.

Datasets

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

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

[2] www.imdb.com

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 2 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 (friend of a friend) 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 reply. 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 [3] 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 [5] community detection algorithm based on modularity, then compute community distances and the community centrality. As an illustration of the resulting communities, we show the common words used in the communities with high community centrality in Figure 3. From the frequent keywords, we easily identify these communities as mainstream computer science fields.

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 [12] 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.

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

AI, Logic Data, Web, Multimedia

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

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 measures 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 [7]. Furthermore, many factors other than an actor's prominence (such as marketing and herd behavior [19]) play a significant role in a movie's success at the box office. Movie budgets, on the other hand, are a measure of how strongly the movie industry believes that a particular 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:

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

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.

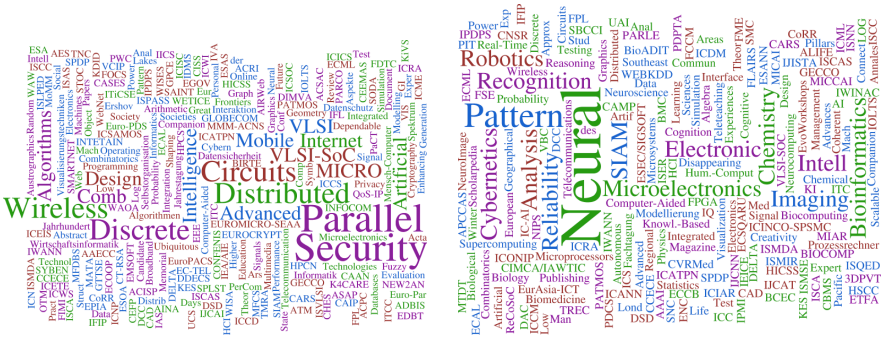


Fig. 3 Common words in peripheral communities in DBLP.

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 [25] 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 $\text{freq}(A)$, and compute $\text{freq}(A_i)/\text{freq}(A_j)$ for all the pairs we study. The results are in the range between 0.0004

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

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

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

H-Index	
L→G	G→L
ccc***	**ccc*****
lcc*	**deg*****
deg***	

TC-10	
L→G	G→L
lcc***	**size*****
deg*	**deg****

(a) DBLP

audience	
L→G	G→L
1***	1***
lcc***	**cc*****
cc*	

(b) boards.ie

budget	
L→G	G→L
1***	**1*****
ccc***	**ccc*****
cc***	**cc*****

gross	
L→G	G→L
1***	**ccc*****
cbc***	**lcc*****
cc**	**cc*****

rating	
L→G	G→L
1***	**1*****
size***	**size****

(c) IMDB(2000s)

Table 4 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 [5] 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.

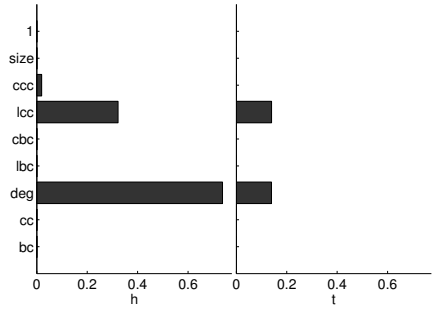


Fig. 4 Weights for predicting H-Index (h) and TC-10 (t) in DBLP data. Light bars indicate negative weights and dark bars indicate positive weights (L→G).

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.

DBLP. The most predictive features are shown in Table 4 and the weights are shown in Figure 4. We see that degree is by far the most predictive feature in this dataset. The more

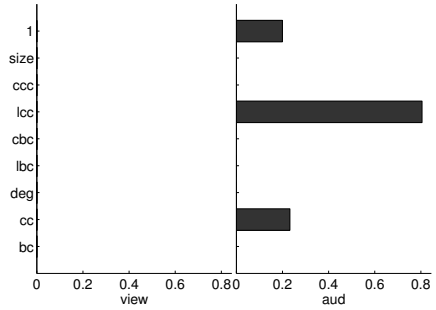


Fig. 5 Weights for predicting views and audience in boards.ie data. Light bars indicate negative weights and dark bars indicate positive weights (L→G).

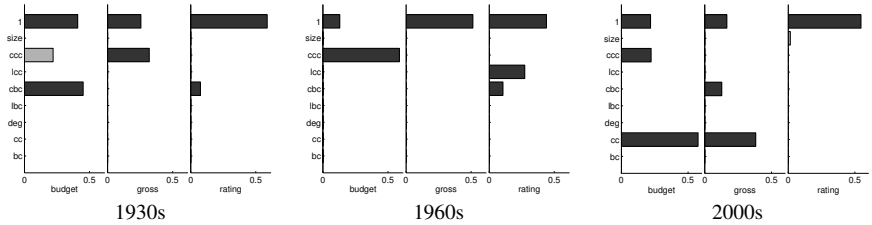


Fig. 6 Weights for predicting budget, gross and rating in IMDB data from 1930s, 1960s, 2000s. Light bars indicate negative weights and dark bars indicate positive weights (L→G).

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 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 4 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 4 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.

Summary. 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.

Impact of community detection method.

In this section, we investigate the robustness of the results to the choice of community detection algorithm. We have applied the Walktrap algorithm [17] 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 5. 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		gross		rating	
L→G	G→L	L→G	G→L	L→G	G→L
$\mathbf{1}^{***}$	$\mathbf{1}^{***}$	$\mathbf{1}^{***}$	$\mathbf{ccc}^{***}$	$\mathbf{1}^{***}$	$\mathbf{1}^{***}$
$\mathbf{ccc}^{***}$	$\mathbf{ccc}^{***}$	$\mathbf{ccc}^{***}$	$\mathbf{lcc}^{***}$		
$\mathbf{size}^{***}$	$\mathbf{cc}^{***}$	$\mathbf{cc}^{**}$	$\mathbf{cc}^{***}$		
$\mathbf{cbc}^{*}$					
$\mathbf{cc}^{***}$					

Table 5 The most predictive centrality features for the IMDB 2000s dataset with communities detected by the WalkTrap [17] algorithm.

Prominence measures for different partitions of data

In previous sections, we have shown that community based centrality measures **lcc**, **ccc**, **lbc**, **cbc** provide novel information on top of the classical centrality measures that compute the

same information at the global level. In this section, we conduct an in depth study of the same set of centrality measures. Our aim is to answer the following questions:

- Are the results robust to the selection of the data points?
- Are the results sensitive to the inclusion of specific centrality methods?
- Are the results robust for ordinal evaluation of fit?

In previous sections, we have seen that local and community centrality methods play a role in various networks. For example, in DBLP, the position of the community is important for achieving high H-Index values. We have argued that being in a central community, a foundational field, is important for obtaining high H-Index when we consider the DBLP dataset as a whole. Similarly, we have seen that local closeness centrality was important for obtaining high TC-10 values. However, the networks we study have highly skewed properties: few individuals with high connectivity, and many individuals with very low connectivity. The question we would like to answer in this section is the following. Are our local and community based centrality measures are useful for predicting prominence values at the full network level or in predicting a specific partition of data with less skewed properties? We investigate this by partitioning the data into two main groups: outliers and others. We repeat our analysis in each partition and check whether our new methods provide additional insight even when considering these partitions.

To illustrate this, suppose we divide the DBLP dataset into two groups: highly prolific researchers with many papers (HPub) and the other researchers with relatively few publications when compared to the prolific researchers (LPub). Obviously, LPub contains most of the people in the DBLP dataset. For this group, there is fairly small variation in the H-Index values. HPub contains a small number of people but with very distinctive H-Index values. Instead of combining these two groups together, we can tune the prediction to each group separately. We would like to find out what network measure is best at predicting that an individual with (relatively) few publications will achieve a (relatively) high H-Index. Similarly, we would like to check whether for individuals with high number of publications, their H-index is best predicted by the number of their publications or whether other network measures play a role. In addition, we would like to see if local and community based centrality methods are relevant for these individual partitions.

Hence, to study robustsness of local and community based to centrality to different partitions, we introduce a new study. We divide the network based on two separate criteria:

- degree of individuals in the network, and
- the number of publications (or movies) the individuals have.

For each partition, we consider two groups: low and high value for the given criteria. For degree, we partition by the mean: high partition is the actors with degree higher than the mean and the low partition is the remaining actors. For number of publications, we put only the outliers in the high partition, i.e. actors with publications three standard deviations above the mean. The rest are placed in the low partition. We run FSS as before using all the features listed in Table 1. To make results comparable, we add the partition criteria to the control variables, i.e. number of publications or movies in addition to degree that we used earlier.

To test robustness of results to the inclusion of specific centrality measures, we first compute a best fit linear model using FSS. Then, we recompute FSS by requiring a specific centrality measure to be present in the model. As FSS is a heuristic method, it chooses randomly among measures of similar predictive power. It is also possible that the choice of

Measure	Description	Dataset
kth	kendall-tau with respect to H-Index	DBLP
ktt	kendall-tau of TC-10	DBLP
ktb	kendall-tau of budget	IMDB
ktg	kendall-tau of gross	IMDB
ktr	kendall-tau of rating	IMDB

Table 6 Abbreviations used to denote Kendall-tau based rank correlation values with respect to a given ground truth and an ordering given by an algorithm. Kendall-tau values range between (-1,1), with 1 indicating the highest level of correlation.

one measure may result in another being disregarded. So, to be able to fully explore which combinations of centrality measures best predict a given ground truth, we force the FSS algorithm to consider these different combinations. However, it is likely that some of these combinations are inferior in their fit when compared with the FSS with no restrictions. As a result, we also compute how good a linear model is independently.

Finally, we introduce an ordinal evaluation of fitness to measure the robustness of the results. It is often hard to judge the fitness of an algorithm or a linear model on truly quantitative terms. For example, if a researcher's H-Index is 10% higher than another's, this does not necessarily mean that they are 10% more important. Similarly, a rating of 5 for a movie may indicate that it is twice as good a rating of 4. Since we do not know the actual quantitative values of ground truth measurements, it is often better to compare the ordering induced by a given linear model and the ordering induced by a ground truth value. To achieve this, we use the Kendall-tau rank correlation measure which is computed between two ordered sets a and b over the same set of objects and is given by $kt(a,b) = (X - Y)/Z$ where X (and Y) are the total number of pairs that agree (and disagree resp.) in their ordering with respect to a and b, and Z is the total number of pairs that are compared. Ties require special attention because breaking a tie is not an agreement or a disagreement. To this end, we introduce a penalty of half disagreement in such cases whenever two pairs are tied in one ranked list and are not tied in the other ranked list. Kendall-tau values range between (-1,1), where 1 stands for complete agreement of orderings, i.e. highest correlation between two ordered sets. A measure of -1 is obtained when one ranking is the reverse of the other.

We run FSS using the scores obtained by different centrality measures and control variables. We find the best linear model that predicts the scores of a ground truth value using leave one out cross correlation as before. We then evaluate the resulting linear model by comparing the ordering imposed by this model with the ordering from ground truth value. We use the following shorthand given in Table 6 to show the Kendall-tau based correlation values computed between a given linear model and a specific ground truth value.

Note that we obtain a linear model using scores, not ranking. It is also possible to directly learn a linear model using rankings. We will report on this approach at the end of this section, but concentrate mainly on models obtained with respect to given scores.

DBLP

In this section, we present linear models obtained for different partitions of the DBLP data set side by side for each different ground truth measure. We first list the FSS model run with no restrictions. Then, we list the models requiring different global centrality measures. The statistics for the DBLP partitions are listed in Table 7.

Partition	Cutoff point	Size of the low group	Size of the high group
Degree	7	362,878	118,585
Number of publications (**pubs**)	106	50,073	985

Table 7 The size of the partitions for the DBLP dataset. Note that for the partition for number of publications, we have excluded all authors with less than 10 papers as we cannot obtain a reliable H-Index value for them. The cutoff point is obtained for this dataset of 10+ more papers.

Partitioning by degree. The partitions by degree are shown in Table 8. The linear models predicting H-Index are shown in Table 8(a), and the models for TC-10 are given in Table 8(b). For each linear model, we report on the Kendall-tau performance with respect to all the ground truth values, not the specific one the model is trained on.

First, we note that requiring a specific centrality measures in FSS does not have a significant impact in performance for H-Index. All centrality measures provide more or less the same performance for both low and high degree. However, there is a performance penalty for choosing **cc** for low degree authors. It is easy to see this for individuals who are attached to a high degree person. Even though their closeness centrality is higher as a result of being connected to such a person, this does not mean that they are more prominent in the network. Given that the constant factor is the best predictor and present in almost all the models for low degree, we can conclude that centrality measures are not particularly useful in this case.

For high degree, number of publications, **deg** and either **ccc** or **cc** are useful for predicting H-Index. As **cc** replaces **ccc** only when it is required by FSS, we can argue that the centrality of one's community is the driver for the performance gain in this partition. This also follows with our earlier finding that community centrality is useful for H-Index, but it appears only for high degree actors.

A similar pattern is true for TC-10 as well. Prediction is very hard for low degree actors and roughly the same when requiring different global centrality measures. The number of publications is the most significant predictor of high TC-10 values.

Partitioning by number of publications. The partitions by the number of publications are shown in Table 9. The linear models predicting H-Index are shown in Table 9(a), and the models for TC-10 are given in Table 9(b). In contrast with the previous partition, the partition for few publications include most of the actors in the network while the high partition has onyl the outliers.

In this case, a different picture emerges when compared to degree partitioning. First of all, the prediction for individuals with low number of publications is dependent on various centrality measures. Furthermore, the prediction with respect to H-Index and TC-10 is better for individuals with few publications than those with many publications. This could be due to the size of the dataset. For individuals with few publications, all models perform roughly the same with the exception of **cc** for TC-10. Degree (**deg**) seems to be the most important feature for both ground truth values.

For individuals with many publications, **cc** for H-Index and **lcc** for TC-10 are important. Note that we have seen that **lcc** was predictive for TC-10 in global ordering as well. This appears to be a particularly useful factor for differentiating among individuals with high number of publications.

For individuals with few publications (when compared to the rest of the network), the actual number of publications (**pubs**) is important. However, for individuals with many publications, the actual number of publications is no longer important. The network location is more crucial. The publication count becomes noisy for the outliers and prominence is hard

Case	Features & Weights		kth	ktt	Features & Weights		kth	ktt
Full	**1*****	.04	**.22**	.0^{+}	**pubs***** **ccc***** **deg***	.72 .03 .27	**.36**	.23
Require **deg**	**deg***** **1*****	.0^{+} .04	**.22**	.0^{+}	**deg***** **pubs***** **ccc*****	.27 .72 .03	.33	.25
Require **cc**	**cc***** **1*****	-.03 .06	.05	.03	**cc***** **pubs***** **deg***	.04 .68 .27	**.36**	.23
Require **bc**	1***	.04	**.22**	.0^{+}	**bc***** **pubs***** **ccc***** **deg***	-.36 .78 .03 .27	**.36**	.23
	Authors with low degree				Authors with high degree			

(a) H-Index

Case	Features & Weights		kth	ktt	Features & Weights		kth	ktt
Full	**1****	.02	.19	.0^{+}	**pubs*****	.22	.37	**.27**
Require **deg**	**deg***	.02	- .19	.0^{+}	**deg*****	.20	.33	.25
Require **cc**	**cc****	.03	-.01	-.01	**cc***** **pubs***	.01 .16	.37	**.27**
Require **bc**	**1****	.02	.22	.0^{+}	**bc***** **pubs*****	-.09 .23	.37	**.27**
	Authors with low degree				Authors with high degree			

(b) TC-10

Table 8 The most predictive features in DBLP partitions by degree using all features shown in Table 1 plus number of publications. In all runs except for the full case, we require a specific algorithm to be used in the linear model. **1** represents the constant factor. For each factor, we use * for significance at 10%, ** for significance at 5%, and *** for significance at 1%.

to guess with measures based on their social status. Other factors should be considered such as social influence or herd effects for citation and acceptance rate of their publications.

Summary of DBLP Results. Degree overall is the most important centrality measure in DBLP regardless of how it is partitioned. Closeness centrality is not useful for ordering low degree and low publication individuals in this network. Prediction of prominence is hard with respect to TC-10 for such individuals as well.

Local centrality is important for TC-10 for actors with many publications as it is more important to do discipline specific work to get high number of citations.

Community centrality is important for H-Index for high degree actors. When one has many collaborators, it is important to be in a foundational field to get a high H-Index value.

IMDB

We apply the same partitions to IMDB. For IMDB, we add a feature based on the number of movies, instead of the number of publications. We compute the Kendall-tau based correlation measures for budget, gross and ratings (`ktb`, `ktg` and `ktr` respectively). The statistics for the IMDB partitions are listed in Table 10.

Case	Features & Weights	kth	ktt	Features & Weights	kth	ktt
Full	**deg***** .33 **pubs***** .19 **cc***** .06	**.33**	.22	**cc***** .29 **deg*** .40	.17	.17
Require **deg**	**deg***** .33 **pubs***** .19 **cc***** .06	**.33**	.22	**deg***** .48 **lcc***** .29	**.18**	.18
Require **cc**	**cc***** .06 **pubs***** .19 **deg***** .33	.32	.22	**cc***** .29 **deg*** .40	.17	.17
Require **bc**	**bc***** -.16 **deg***** .34 **pubs***** .20 **cc***** .06	.32	.22	**bc***** .01 **cc***** .29 **deg*** .40	.17	.17
	Authors with few publications			Authors with many publications		

(a) H-Index

Case	Features & Weights	kth	ktt	Features & Weights	kth	ktt
Full	**deg***** .18	.30	**.24**	**lcc***** .13	.21	**.19**
Require **deg**	**deg***** .18	.30	**.24**	**deg**** .25	.18	.14
Require **cc**	**cc***** .03	.19	.17	**cc***** .11	.21	.17
Require **bc**	**bc*** .05 **deg***** .17	.30	.23	**bc***** .02 **lcc***** .12	.21	**.19**
	Authors with few publications			Authors with many publications		

(b) TC-10

Table 9 The most predictive features in DBLP for partitioning by the number of publications using all features shown in Table 1, plus the number of publications. In all runs except for the full case, we require a specific algorithm to be used in the linear model. **1** represents the constant factor. For each factor, we use * for significance at 10%, ** for significance at 5%, and *** for significance at 1%.

Partition	Cutoff point	Size of the low group	Size of the high group
Degree	5	32,557	9,020
Number of movies (**movies**)	2.5-3	30,535	2,022

Table 10 The size of the partitions for the IMDB dataset. Note that for the partition for number of movies, we considered the mean number of movies for actors and directors separately and placed them in high/low groups accordingly.

Partitioning by degree. The results of degree based partitioning are given in Table 11. Models are trained with respect to budget in part (a), ratings in part (b), and gross in part (c). We use on the 2000s in this prediction.

Overall, budget prediction offers the best prediction rates. We see a few interesting patterns. For low degree actors, degree is not important, but **cc** and **ccc** are both important. For high degree actors, while **cc**, **ccc** are both important, it appears **bc** is crucial for ordering individuals. Note that models incorporating **bc** are not optimal for scores, as **bc** is not picked by score based regression unless it is explicitly required. When it is picked, the weight for **bc** is negative. To find who will be in high budget movies, we need to look at actors with low betweenness value, but high global and community closeness centrality values. For this model, our prediction reaches 50%, a particularly high number. For all other models for high degree actors, there is no rank correlation for budget. In fact, budget is the measure that is most dependent on the network as it shows who the industry trusts to invest money on.

Ratings are very hard to predict and most models perform quite poorly. For gross, prediction is not sensitive to the selection of any particular centrality measure. Global closeness centrality (**cc**) is important for both low and high degree actors, while community closeness

Case	Features & Weights	ktb	ktr	ktg	Features & Weights	ktb	ktr	ktg
Full	**cc***** .51 **1***** .22 **ccc**** .23	**.36**	-.06	.20	**cc***** .65 **1***** .24 **ccc***** .19 **movies*** -.96 **deg*** .77	.0^{-}	-.01	.02
Require **deg**	**deg***** .09 **cc***** .52 **1***** .18 **ccc*** .22 **movies*** -.50	.29	-.05	.19	**deg***** .77 **cc***** .65 **1***** .24 **ccc***** .19 **movies*** -.96	.0^{-}	-.01	.02
Require **cc**	**cc***** .51 **1***** .22 **ccc**** .23	**.36**	-.06	.20	**cc***** .65 **1***** .24 **ccc***** .19 **movies*** -.96 **deg*** .77	.0^{-}	-.01	.02
Require **bc**	**bc*** -.46 **cc***** .55 **1***** .27 **cbc**** .15	**.36**	-.05	.20	**bc***** -.43 **cc***** .63 **1***** .21 **ccc**** .20 **deg*** -.02	**.49**	-.02	.25
	Actors with low degree				Actors with high degree			

(a) Budget

Case	Features & Weights	ktb	ktr	ktg	Features & Weights	ktb	ktr	ktg
Full	**1***** .55 **size*** -.02	-.24	**.04**	-.16	**1***** .54 **size**** -.01	-.41	**.02**	-.24
Require **deg**	**deg***** -.03 **1***** .56 **size*** -.02	-.24	**.04**	-.18	**deg***** .07 **1***** .54 **size**** -.02	-.41	**.02**	-.23
Require **cc**	**cc***** -.08 **1***** .59 **size*** .0^{+}	.23	-.03	.13	**cc***** -.10 **1***** .59 **size**** .01	.40	-.02	.20
Require **bc**	**bc***** -.02 **1***** .55 **size***** -.02	-.24	**.04**	-.16	**bc***** -.18 **1***** .55 **size**** -.01	-.39	**.02**	-.23
	Actors with low degree				Actors with high degree			

(b) Ratings

Case	Features & Weights	ktb	ktr	ktg	Features & Weights	ktb	ktr	ktg
Full	**cc***** .69	.18	.01	**.20**	**cc***** .51 **ccc***** .21	.21	.06	.28
Require **deg**	**deg***** .15 **cc***** .57	.18	.01	**.20**	**deg***** -.24 **cc***** .64 **cbc**** .14	.23	.07	.29
Require **cc**	**cc***** .69	.18	.01	**.20**	**cc***** .51 **ccc**** .21	.21	.06	.28
Require **bc**	**cc***** .69	.18	.01	**.20**	**bc***** -.75 **cc***** .65 **cbc**** .12	.24	.08	**.30**
	Actors with low degree				Actors with high degree			

(c) Gross

Table 11 The most predictive features in IMDB for partitioning by the degree of the actors using all features shown in Table 1. In all runs except for the full case, we require a specific algorithm to be used in the linear model. **1** represents the constant factor. For each factor, we use * for significance at 10%, ** for significance at 5%, and *** for significance at 1%.

centrality (**ccc**) is important for high degree actors. Prediction of ratings is quite poor in both partitions and inclusion **cc** hurts prediction slightly.

Partitioning by number of movies. The results of partitioning based on the number of movies are given in Table 12. Models are trained with respect to budget in part (a), ratings in part (b), and gross in part (c). We use on the 2000s in this prediction.

We note that global and community centrality (**cc**, **ccc**) remain very important for predicting budget for both actors with few and many movies. The prediction is in the range 40-50% for both partitions. The only centrality measure that appears to impact prediction accuracy very negatively is degree in both partitions.

Ratings is again similarly hard to predict. Including betweenness centrality (**bc**) with a negative weight results in a small improvement. Similarly, for gross, betweenness centrality plays a central role with a negative weight. Only when **bc** is included, we get good prediction accuracy for both partitions together with **cc** and **ccc**.

Summary of IMDB Results. For IMDB, the partitions behave differently. Unlike DBLP, requiring degree centrality often hurts performance. Degree is not useful to predicting all the measures we consider here in almost all partitions. However, global and community closeness centrality are very important together. It is not who you know, it is where you and your community are in the network.

The Kendall-tau correlation for the models predicting budget are quite high, reaching 0.5. Budget is the one feature in our dataset that is highly dependent on the network. Having low betweenness centrality is crucial in particular for budget. Actors that star in movies from many different communities (i.e. have high betweenness) do not star in big budget movies. Such high-betweenness actors are in some sense "generic" actors. One thing we point out is that, in our network, we only consider the top three actors for each movie according to the billing order. Often the movie database list all actors that star in them no matter how small their role is. We concentrate on the top three as the actors with the most important role in that movie. These actors also tend to be the top paid actors. Hence, it is likely that actors with high betweenness are often in high budget movies, but not as the "star". They most likely have secondary and supporting roles.

Similar to the results for DBLP, local and community centrality measures are relevant for different partitions of data as we as the whole network. As a result, we conclude that these new measures provide a new way to analyze networks and its different partitions.

As we discussed earlier, our results so far concentrated on using scores to find an optimal linear model. We then evaluated these models based on the rank ordering that they provide. It is also possible to directly predict the rank ordering using a heuristic based on linear programming. In this method, the ordering of individuals based on rank are translated into a set of constraints. The linear programming is then used to find a linear model that satisfies the largest number of these constraints possible [5]. Note that due to the high computation cost of this method, we do not use cross validation.

The caveat of using this method is that ranks alone provide less information than scores. By supressing the score information we are reducing both noise and signal. As a result, it is not guaranteed that this method will result in better prediction overall for the Kendall-tau ordering. In fact, the Kendall-tau correlation of linear models found using this method and a given ground truth are always inferior to the ones we reported earlier in this paper. This is

[5] The implementation of this algorithm is given in `http://www.cs.rpi.edu/~magdon/LFDlabpublic.html/software.html?`

Case	Features & Weights	ktb	ktr	ktg	Features & Weights	ktb	ktr	ktg
Full	**cc***** .54 **1***** .21 **ccc***** .22 **deg*** .03	**.42**	.06	.20	**cc***** .63 **ccc*** .24 **1*** .14	**.54**	.02	.27
Require **deg**	**deg***** .60 **cc***** .57 **1***** .18 **ccc***** .20 **movies***** -.39	.09	.0^{+}	.12	**deg***** .58 **cc***** .59 **ccc*** .22 **1*** .21 **movies*** -.67	-.10	-.01	.01
Require **cc**	**cc***** .54 **1***** . 21 **ccc***** .22 **deg*** .03	**.42**	-.06	.20	**cc***** .63 **ccc*** .24 **1*** .14	**.54**	.02	.27
Require **bc**	**bc***** -.33 **cc***** .55 **1***** .21 **ccc***** .21 **deg*** .13	**.42**	-.06	.20	**bc***** -.43 **cc***** 1.08	.52	-.01	.27
	Actors with few movies				Actors with many movies			

(a) Budget

Case	Features & Weights	ktb	ktr	ktg	Features & Weights	ktb	ktr	ktg
Full	**1***** .55 **size***** -.02	-.29	**.04**	.19	**1***** .55 **size***** -.01	-.47	-.02	-.22
Require **deg**	**deg***** -.02 **1***** .55 **size*** -.02	-.30	**.04**	-.21	**deg***** -.03 **1***** .55 **size*** .0^{-}	-.48	-.02	-.23
Require **cc**	**cc***** -.09 **1***** .59 **size*** .0^{+}	.29	-.03	.16	**cc***** -.13 **1***** .66 **cbc*** .45 **size**** -.44	-.50	-.02	-.31
Require **bc**	**bc***** -.09 **1***** .55 **size*** -.02	-.29	**.04**	-.20	**bc***** -.30 **1***** .56 **size*** .0^{+}	.49	**.06**	.29
	Actors with few movies				Actors with many movies			

(b) Ratings

Case	Features & Weights	ktb	ktr	ktg	Features & Weights	ktb	ktr	ktg
Full	**cc***** .37 **ccc*** .62 **size*** -.34	-.20	-.05	-.16	**cc***** .55 **cbc***** .17	.28	.11	.31
Require **deg**	**deg***** .61 **cc***** .53 **ccc*** .13 **movies*** -.47	.16	.10	.05	**deg***** -.24 **cc***** .63 **cbc*** .17	.29	.11	.31
Require **cc**	**cc***** .37 **ccc*** .62 **size*** -.34	-.20	-.05	-.16	**cc***** .55 **cbc*** .17	.29	.11	.31
Require **bc**	**bc***** -.47 **cc***** .57 **ccc*** .13	.18	-.01	**.24**	**bc***** -.60 **cc***** .62 **cbc*** .14	.32	.13	**.35**
	Actors with few movies				Actors with many movies			

(c) Gross

Table 12 The most predictive features in IMDB for partitioning by the number of movies using all features shown in Table 1. In all runs except for the full case, we require a specific algorithm to be used in the linear model. **1** represents the constant factor. For each factor, we use * for significance at 10%, ** for significance at 5%, and *** for significance at 1%.

true for the full network, as well as its many partitions. As a result, we do not provide the models found by this method here.

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. We have also studied the robustness of our measures by looking at different partitions of data, for different linear models and for ordinal evaluation of the given models.

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 anyone in the network is likely to be familiar 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. To make a high impact with high number of citations, it is important to be central in a specific community. In essence, people who are deeply embedded in a specific community tend to get high citations for their work. Local centrality becomes important to distinguish between outliers in a specific community.

Our deconstruction provides us with an enhanced vocabulary when studying prominence of individuals in a network and the factors that contribute to it. Based on the underlying measure of prominence, we can study the factors that contribute to prominence in the network both for the whole network or specific partitions of it. Local centrality suggests a prominence measure that is based on local processes. It is a good to tool study micro processes like paper citations at the individual paper level, not at the H-Index level. Community centrality suggests network wide processes like diffusion of information between different communities or distribution of resources in the network. 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 [7]. Often more than one of these processes are active in a network when determining prominence. As a result, our measures provide a way to better tune the structural analysis of networks at various levels of granularity and understand the factors contributing to prominence.

Many interesting problems remain. We would like to use our methods to evaluate different variations of centrality measures proposed in the literature and analyze to which degree they capture the local, global and community based information. As we have seen, betweenness becomes an important measures for ranking high degree actors. We would like to inves-

tigate this further and understand community betweenness better. The local and community based 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 [1,13] 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 other networks and gain further insight into prominence in these networks. We can also further study local and community versions of other centrality methods. For example, local and network degree would differ significantly for researchers conducting interdisciplinary research.

We also can extend the community centrality measure to overlapping communities which constitute a more natural way to group individuals in social networks [10, 15, 18]. Another interesting study is the evaluation of community detection methods by considering how well local centrality measures perform for different prominence measures using communities discovered by different algorithms [23]. If communities are meaningful units within the network, then local centrality in these communities must be a useful measure. Hence, we can target prominence measures that are particularly dependent at local processes to study how well community detection algorithms work.

Acknowledgments

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. Xiahui Lu is supported by DARPA SMISC program via a subcontract to RPI from Sentametrix. 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. S. Adalı, X. Lu, and M. Magdon-Ismail. Deconstructing centrality: thinking locally and ranking globally in networks. In *Proceedings of 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, 2013.
3. L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.
4. S. P. Borgatti and M. G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466 – 484, 2006.
5. A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys Rev E*, 70(6):066111+, 2004.
6. T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 3rd edition, 2009.
7. A. De Vany. *Hollywood economics: How extreme uncertainty shapes the film industry*. London: Routledge, 2004.
8. M. Everett and S. P. Borgatti. *Extending Centrality*. Cambridge University Press, 2005.
9. L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.
10. I. F. G. Palla, I. Derenyi and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, pages 814–818, 2005.
11. R. Guimera and L. Amaral. Modeling the world-wide airport network. *Eur. Phys. J. B*, 38:381–385, MAR 2004.

12. J. Hirsch. An index to quantify an individual's scientific research output. *Proc. of the National Academy of Sciences*, 46:16569–16572, 2005.
13. N. Kourtellis, T. Alahakoon, R. Simha, A. Iamnitchi, and R. Tripathi. Identifying high betweenness centrality nodes in large social networks. *Social Network Analysis and Mining*, 3(4):899–914, 2013.
14. A. Langville and C. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1:335–380, 2005.
15. M. Magdon-Ismail and J. Purnell. Ssde-cluster: Fast overlapping clustering of networks using sampled spectral distance embedding and gmms. In *IEEE International Conference on Social Computing*, 2011.
16. 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.
17. P. Pons and M. Latapy. Computing communities in large networks using random walks. In *Proc. 20th Comp. and Inf. Sc.*, pages 284–293, 2005.
18. B. Rees and K. Gallagher. Overlapping community detection using a community optimized graph swarm. *Social Network Analysis and Mining*, 2(4):405–417, 2012.
19. 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.
20. J. Scott. *Social network analysis*. SAGE Publications Limited, 2012.
21. K. Stephenson and M. Zelen. Rethinking centrality: Methods and examples. *Social Networks*, 11(1):1–37, Mar. 1989.
22. 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.
23. M. Vasudevan and N. Deo. Efficient community identification in complex networks. *Social Network Analysis and Mining*, 2(4):345–359, 2012.
24. S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
25. 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.
26. J. Zhao, G.-H. Ding, L. Tao, H. Yu, Z.-H. Yu, J.-H. Luo, Z.-W. Cao, and Y.-X. Li. Modular co-evolution of metabolic networks. *BMC Bioinformatics*, 2007.