AFRL-RI-RS-TR-2015-077

ANALYZING EVOLVING SOCIAL NETWORKS 2 (EVOLVE2)

UNIVERSITY OF SOUTHERN CALIFORNIA

APRIL 2015

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th ABW, Wright-Patterson AFB Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2015-077 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/ S /
TODD WASKIEWICZ
Work Unit Manager

/ S /
MICHAEL J. WESSING
Deputy Chief, Information Intelligence
Systems & Analysis Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
APRIL 2015	FINAL TECHNICAL REPORT	JUN 2012 – OCT 2014

4. TITLE AND SUBTITLE

ANALYZING EVOLVING SOCIAL NETWORKS 2 (EVOLVE2)

5a. CONTRACT NUMBER
FA8750-12-2-0186

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
63788F, 62788F

6. AUTHOR(S)

Kristina Lerman

5d. PROJECT NUMBER
E3NA

5e. TASK NUMBER
DS

5f. WORK UNIT NUMBER
N2

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Southern California
3720 S. Flower Street, Third Floor
Los Angeles, CA 90089-0001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIEA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2015-077

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. PA# 88ABW-2015-1769
Date Cleared: 8 APR 2015

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Current social network analytic methods analyze a static aggregate graph, which provides a limited view of the structure and behavior of real world social networks. Real world networks are dynamic: they evolve over time as new connections form between individuals, and networks themselves act as a substrate for the flow of information and influence. Ignoring dynamics can produce a distorted, and even wrong, view of who the important individuals are in a social network, what is the nature and strength of the connections between them, and what are the communities of similar or similarly behaving individuals. The erroneous conclusion reached by static network analysis will waste analysts' time and resources. For these reasons, we developed network analysis methods that directly incorporate time. The research had two major threads: -Understand how networks evolve over time, and how changes in topology affect evolution of influence and groups -Understand the impact of dynamics and network flows on the measurement of the network structure

15. SUBJECT TERMS

Dynamic social network analysis, bonacich centrality, community detection, heterogeneous social networks

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**TODD WASKIEWICZ**
U	U	U	UU	24	19b. TELEPHONE NUMBER (*Include area code*) **NA**

Contents

List of Figures

List of Tables

1 OBJECTIVES

Current social network analytic methods analyze a static aggregate graph, which at provides a limited view of the structure and behavior of real-world social networks. Real-world networks are dynamic: they evolve over time as new connections form between individuals, and networks themselves act as a substrate for the flow of information and influence. Ignoring dynamics can produce a distorted, and even wrong, view of who the important individuals are in a social network, what is the nature and strength of the connections between them, and what are the communities of similar or similarly-behaving individuals. The erroneous conclusions reached by static network analysis will waste analysts' time and resources.

For these reasons, we proposed to develop network analysis methods that directly incorporate *time*. The research had two major threads:

- Understand how networks evolve over time, and how changes in topology affect evolution of influence and groups
- Understand the impact of dynamics and network flows on the measurement of network structure

Progress in dynamic network analysis was hampered by scarcity of large-scale network data sets with fine-grained temporal resolution. We mainly worked with two dynamic network data sets: citations data, representing citations between physics articles over a period of 100 years, and online social network data we collected from social media sites such as Digg and Twitter. While some of the models and observations are limited to these particular network data sets, we believe that the methods and approaches we developed using these data sets will generalize to other dynamic networks.

Below we describe our technical approach to address these questions and significant accomplishments.

2 DYNAMIC NETWORK ANALYSIS

2.1 Measuring Influence in Dynamic Networks

Centrality measures a node's importance or influence in a network. Over the years a variety of measures have been proposed for node centrality, including degree centrality, Katz status score [1], alpha-centrality [2], eigenvector [3] and betweenness centrality [4], and variants based on random walk, such as PageRank [5]. Consider, specifically, alpha-centrality as defined by Bonacich, which measures the total number of paths of any length between two nodes i and j, with longer paths contributing less to the centrality than shorter paths. Let A be the adjacency matrix of a network, such that $A_{ij} = 1$ if an edge exists from i to j and $A_{ij} = 0$ otherwise. The alpha-centrality matrix is given by:

$$C(\alpha) = A + \alpha A^2 + \alpha^2 A^3 + ...$$

where α is the attenuation factor along an edge. This parameters sets the length scale of interactions. For $\alpha = 0$, alpha-centrality takes into account direct edges only and reduces to degree centrality. As α increases, this becomes a more global measure, taking into account more distant interactions. However, α is bounded by the inverse of the largest eigenvalue of A. As $\alpha \to 1/\lambda_{max}$, alpha-centrality approaches eigenvector centrality [6]. In numerous works, we showed that alpha-centrality is a useful measure of identifying important nodes in a network [7, 8, 9, 10]. In a *dynamic network*, where edges may change over time, the notion of a path must be refined to include time. To this end, we defined dynamic alpha-centrality matrix [11], which considers paths over time-dependent edges in a network.

In addition to dynamic alpha-centrality, we introduced two new measures of centrality for growing networks. The first, effective contagion matrix [12], overcomes the *recency bias* of centrality measures that fail to recognize important new nodes that have not had as much time to accumulate links as their older counterparts. The second approach to dynamic centrality [13] extends the notion of a time-dependent paths introduced in our earlier work to consider all paths, a *cascade*, emanating from a node in a dynamic network. Comparing the size and structure of cascades [14] generated by two nodes enables us to compare them in importance.

Figure 1(a) illustrates our idea. A *seed* (red node) represents an established paper in a field of research. The paper's influence grows over time as new papers cite it and are later cited by other papers, creating a *cascade* of citations that can be traced back to the seed. A *challenger* (blue node) is a paper that advocates a new paradigm. It attracts new citations from papers shown as white nodes with blue background, leaving the *complement cascade* (green nodes)

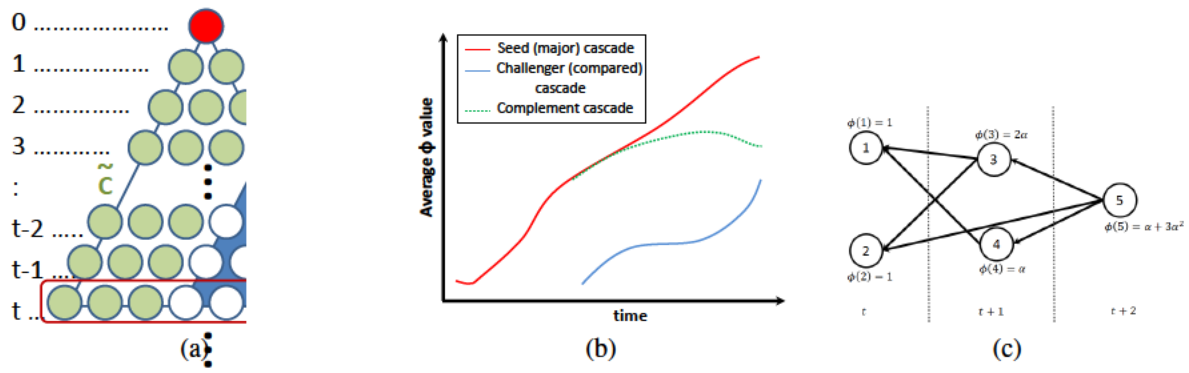


Figure 1: (a) Information disruption by a challenger in an information cascade. The seed of an established paradigm, marked in red, creates a cascade as it is cited by other papers, while a challenger, marked in blue, disrupts the cascade of the seed. (b) Disruption of the cascade of the seed paradigm (red) by the challenger paradigm (blue) can be visualized as the decline of Φ of the complement cascade (green). (c) An example cascade.

containing papers in the cascade of the seed that are not connected to the challenger. When the challenger represents a non-competing idea, though there will be papers that cite both seed and challenger, they will not interfere with the growth of the seed's cascade. In contrast, a transformative challenger will disrupt the growth of influence of the established paper. Without considering the challenger, it may appear that the established paper continues to prosper, as its cascade continues to grow, but subtracting part of the cascade taken over by the challenger will reveal that the growth of the complement cascade (green nodes) slows. In this case, the community's attention shifts to the challenger paradigm.

We briefly summarize the approach. For additional details please see [13]. Given one or more papers $\mathcal{S}$ in a citations graph, a cascade C is a subgraph that contains all citation chains that end at $\mathcal{S}$. The set $\mathcal{S}$ is called the *seed* or *root* of the cascade. The seed indirectly exerts influence on all papers in the cascade, but influence decays with the distance to the seed, just like in alpha-centrality [2]. For a node j in the cascade, the cascade generating function $\phi(j)$ summarizes the structure of the cascade [14], i.e., all existing citation chains. The cascade generating function quantifies the influence of $\mathcal{S}$ on node j, and is defined recursively by

$$\phi(j) := \begin{cases} 1 & \text{if } j \in S \\ \sum_{i \in cite(j)} \alpha\phi(i) & \text{otherwise,} \end{cases} \tag{1}$$

where α is a constant damping factor. Figure 1(c) shows an example cascade and the ϕ values for its nodes. For a paper j published after T time steps (e.g., years) from the publication of the seed, $\phi(j)$ can be written as follows:

$$\phi(j) = \sum_{p=0}^{T} a_p \cdot \alpha^p, \tag{2}$$

where the coefficient a_p is the number of distinct paths of length p from one of the seeds to j. The impact of α is that the smaller the value of α, the higher the penalty against long paths. It is also possible to assign a unique α_{ij} for each link but we found that it is simpler to assign a constant 0.5 for all links to control its impact.

We quantify this decline by the *disruption score* $\delta(\tau)$, which is a function of the time interval of τ given the seed and challenger cascades. Let t_0 be the publication time of the challenger paper,

$$\begin{aligned} \delta(\tau) &:= \sum_{t=t_0}^{t_0+\tau} \log \frac{\Phi_t(C)}{\Phi_t(\widetilde{C})} \\ &= \sum_{t=t_0}^{t_0+\tau} \left(\log \Phi_t(C) - \log \Phi_t(\widetilde{C}) \right). \end{aligned}$$

Table 1: Top ten challengers to the 1957 "Theory of Superconductivity" identified by (a) proposed method and (b) baseline.

Year	Cites	Title
(a) our method: sorted by disruption score		
1958	14	Meissner Effect
1958	307	Random-Phase Approximation ... Superconductivity
1959	40	Evidence for Anisotropy of the Superconducting Energy...
1989	574	Phenomenology of ...Cu-O high-temperature supercon...
1987	368	Antiferromagnetism in La_2CuO_{4-y}
1987	281	Two-dimensional antiferromagnetic quantum ...
1988	149	$Ba_2YCu_3O_7$: Electrodynamics of Crystals ...
1990	156	High-resolution angle-resolved photoemission ...
1988	399	Low-temperature behavior of two-dimensional quantum ...
1995	95	Momentum Dependence of the Superconducting ...
(c) baseline: sorted by citations		
1981	3191	Self-interaction correction to density-functional approx...
1996	3088	Generalized Gradient Approximation Made Simple
1980	2651	Ground State of the Electron Gas by a Stochastic Method
1976	2569	Special points for Brillouin-zone integrations
1996	2387	Efficient iterative schemes for ab initio total-energy...
1990	1951	Soft self-consistent pseudopotentials in a generalized...
1991	1950	Efficient pseudopotentials for plane-wave calculations
1975	1597	Linear methods in band theory
1992	1567	Atoms, molecules, solids, and surfaces:...
1992	1445	Accurate and simple analytic representation...

The disruption score can be visualized as the area between the red and green curves in Figure 1(b) from t_0 to $t_0 + \tau$. The disruption score allows us to identify and measure the impact of the challenger paper.

We applied the disruption score to identify transformative physics articles published by the American Physical Society (APS). Through several case studies we showed that the proposed method is better able to identify successful challengers than alternative baseline that considers the number of citations received by the paper. Further, we demonstrated that our method identifies more relevant challengers than baseline. Moreover, challenger's success is evident early on, allowing for early detection of transformative research.

2.1.1 Case Study.

In 1957 Bardeen, Cooper and Schrieffer published a seminal paper titled "Theory of Superconductivity" which explained the mechanism by which some metals became perfect electrical conductors (i.e., they lost their electrical resistance) at low temperatures. The authors were awarded a Nobel prize for this discovery in 1972. This paper is one of the ten most cited papers in the APS dataset.

Table 1 lists the ten top-ranked challengers identified by our method and the baseline (number of citations). Compared to baseline, our method identifies papers that are relevant to the topic of superconductivity. All ten of the top challengers identified by baseline are papers dealing with calculations of electronic structure of materials, and include other most-cited papers in the APS dataset. While this is a very important topic, it is only peripherally related to superconductivity, in as much as this phenomenon is a result of correlated electron pairs.

The proposed method discovered papers on high temperature superconductivity (HTS). The discovery of HTS was an important development in the study of superconductivity, recognized with a Nobel prize in 1987. Although the original paper announcing the discovery is not in our dataset, presence of several other papers on HTS among the top challengers demonstrates the efficacy of our method to identify disruptive papers. These challengers include "Antiferromagnetism in La_2CuO_{4-y}", "Two-dimensional antiferromagnetic quantum spin-fluid state in La_2CuO_4", "$Ba_2YCu_3O_7$: Electrodynamics of Crystals with High Reflectivity" and "Momentum Dependence of the Superconducting $Sr_2CaCu_2O_8$".

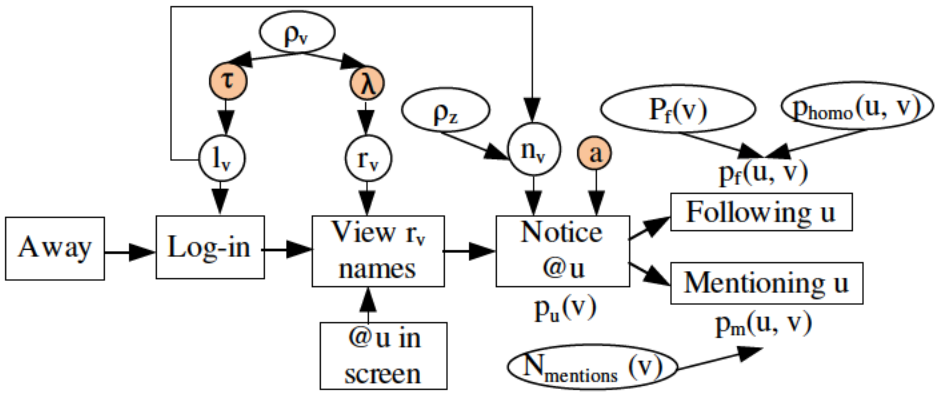


Figure 2: State diagram of a user v's behavior regarding to another user u who is in the screen of v. The ovals represent parameters which are measured, while circles represent parameters to be estimated from data. To simplify the model, only the filled circles which represent the hyper-parameters are to be finally estimated from data.

2.1.2 Relevant Publications.

- Y. hung Huang, C.-N. Hsu, and K. Lerman. Identifying transformative scientific research. In *Proc. of IEEE International Conference on Data Mining*, 2013.
- R. Ghosh and K. Lerman. A framework for quantitative analysis of cascades on networks. In *Proceedings of Web Search and Data Mining Conference (WSDM)*, February 2011.
- R. Ghosh and K. Lerman. Parameterized centrality metric for network analysis. *Physical Review E*, 83(6):066118, June 2011.
- R. Ghosh, T.-T. Kuo, C.-N. Hsu, S.-D. Lin, and K. Lerman. Time-aware ranking in dynamic citation networks. In *COMMPER 2011: Mining Communities and People Recommendations, Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 373 –380, December 2011.
- R. Ghosh and K. Lerman. Predicting influential users in online social networks. In *Proceedings of KDD workshop on Social Network Analysis (SNA-KDD)*, July 2010.
- K. Lerman, R. Ghosh, and J.-h. Kang. Centrality metric for dynamic network analysis. In *Proceedings of KDD workshop on Mining and Learning with Graphs (MLG)*, July 2010.

2.2 Link Prediction in Dynamic Networks

A core task of social network analysis is to predict the formation of new links between nodes. In the context of social media, link prediction serves as the foundation for forecasting the evolution of the follower graph and predicting interactions and the flow of information between users. Previous link prediction methods have generally represented the social network as a graph and leveraged topological and semantic measures of similarity between two nodes to evaluate the probability of link formation. We proposed a link creation mechanism for social media [15] wherein a person v creates a link to person u after seeing u's name on his or her screen. In other words, visibility of a user (name) is a necessary condition for new link formation. This work illustrates our approach of uncovering mechanisms driving behavior in social networks and building models based on these mechanisms.

We proposed a visibility-based model for link prediction, which estimates the probability of a user views another user's name, and used this model to predict evolution of the social media follower graph. Figure 2 shows the process through which a user v from a population views another user u, notices u and then responds to u by deciding to follow u. When a user v visits a social media site to view her stream, she may view a limited portion of her stream, hence seeing only a limited number of screen names. The likelihood v will see u's name in his or her stream depends on the total number of screen names n_v in v's stream, the frequency v visits the social media site, and as well as the average number of names v views per visit. Once v notices u's name with a certain probability $p_u(v)$, v might create new "follow" link.

We created a stochastic model that describes the process of discovering users shown in Fig. 2 and estimated its parameters by a Maximum-Likelihood approach. We applied the model to study dynamics of the Twitter follower graph by predicting new follower links. We used two popular metrics to evaluate the link prediction results: the Area Under ROC Curve (AUC) and mean average precision (MAP). ROC curve is created by plotting the fraction of true positives out of the total actual positives (true positive rate) vs. the fraction of false positives out of the total actual negatives (false positive rate), at various threshold settings. Thus, AUC score evaluates the overall ranking yielded by the algorithms with a larger AUC indicating better link prediction performance.

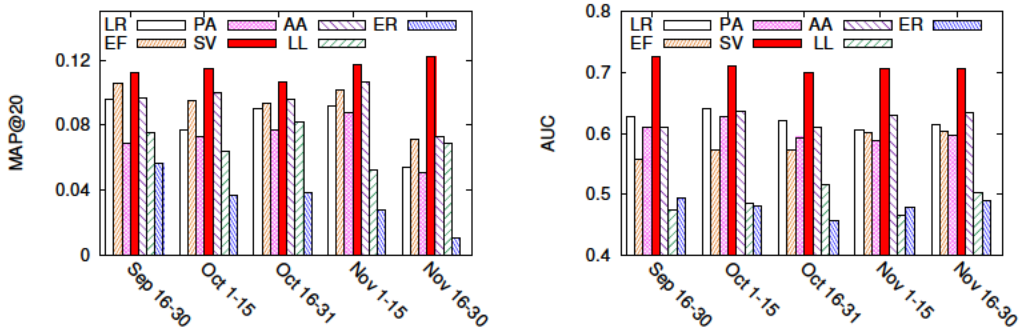


Figure 3: Comparison of different algorithms on the task of predicting follow links using MAP and AUC scores.

We compared the performance of the proposed model to six different baselines, including linear regression (LR), edge rank (ER), preferential attachment (PA), Adamic-Adar (AA) score, latest activity (LL), and exposure frequency (EF). Figure 3 show results of the proposed model to baselines on the task of predicting new follow links. Note that x-axis denotes the testing time interval and all the available data before testing time are for training. The proposed approach SV (solid red color) outperforms all other baselines in terms of both MAP and AUC scores.

2.2.1 Relevant Publications.

- L. Zhu and K. Lerman. A visibility-based model for link prediction in social media. In *Proceedings of the ASE/IEEE Conference on Social Computing*, 2014.
- N. O. Hodas and K. Lerman. How limited visibility and divided attention constrain social contagion. In *ASE/IEEE International Conference on Social Computing*, 2012.

3 DYNAMIC PROCESSES ON NETWORKS

The major focus of this project was to understand how dynamics affected our understanding and measurement of network structure. By dynamics we mean the processes by which nodes in a network interact, e.g., to share money, ideas, information, or influence each other. For example, consider a social network in which people exchange money for goods and services. Each person takes some money and distributes it among neighbors. Since no money is created or destroyed, this type of interaction can be modeled by a random walk. A random walk is a process where at each step a walker chooses a random neighbor of a given node to transition to. While many social processes can be described by random walks (web surfing, used goods exchange, phone calls, etc.) many cannot be thus described. Consider an infectious disease spreading between people. At each point, rather than infecting only one neighbor, a sick person can infect (with some probability) all neighbors. Information, influence and epidemics spread in such non-conservative fashion.

Taking dynamics into account changes our perception of network structure. We use a simple example to illustrate how dynamics affects who the central nodes are and what communities exist in a network. A central node in a money exchange network is one who frequently receives sums of money from others. If this were an illicit exchange network, law enforcement would want to target this central node to disrupt the network's activities. A central node in an infectious disease network, on the other hand, is one who is frequently infected by others. If we had just one vaccine to give, we have to give it to this central node to best stop the spread of the disease. If the same network has both

a virus spreading on it and illicit money exchange taking place, then the node we will want to vaccinate will almost certainly not be the same as the one that would be targeted by law enforcement. Similarly, if we define a community in a money exchange network as a group of nodes who frequently exchange money with each other, it most certainly won't be the same group that frequently infects each other with a virus. Below we clarify and formalize the impact of dynamic processes on network structure.

3.1 Dynamics and Centrality

Existing centrality measures examine link structure of the network to identify key individuals within it. However, as we argued previously, centrality is intimately related to dynamic processes taking place on the network, processes which determine how information, disease or goods flow on the network. For example, by modeling Web surfing as a random walk, PageRank assigns a score to each Web page based on its value in the equilibrium distribution of the random walk. In contrast, central individual in a social network through which a disease is spreading is one who infects most others. Such influential individuals are scored highly by the Katz index or Bonacich's Alpha-Centrality, both of which give the equilibrium distribution of an epidemic process on a network [9, 10].

Now consider information spreading through a population, for instance, by users sending messages or product recommendations to their friends using email or social media. We have demonstrated recently that information spread cannot be modeled as an epidemic diffusion. Instead, cognitive constraints, such as limited attention are important [16]. Attention is the psychological mechanism that controls how we process incoming stimuli and decide what activities to engage in. Actions, such as reading a tweet, browsing a Web page, or reading a science article, require mental effort, and since human brain's capacity for mental effort is limited, so is attention. As a consequence, the more stimuli people have to process, the smaller the probability they will respond to any one stimulus.

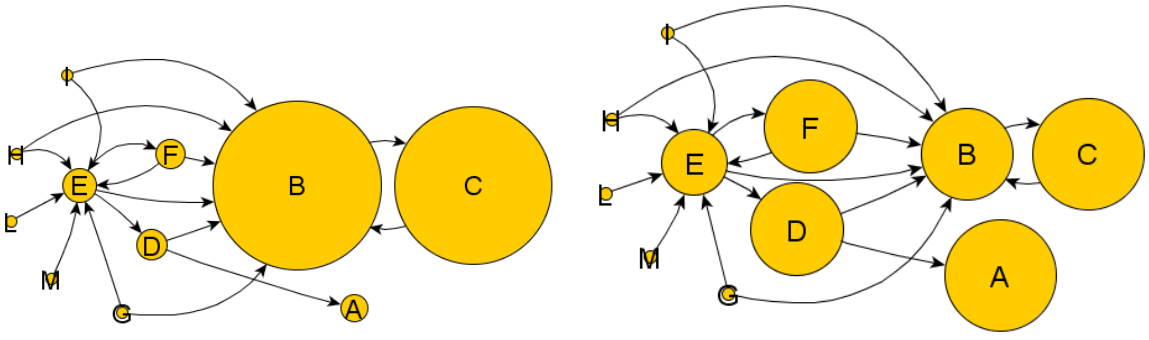
Cognitive constraints the nature of social interactions and therefore, how central nodes are identified. Now a node's capacity to infect others depends not only on how many connections it has but also on who and how many others these nodes are connected to. We have recently introduce a new centrality for social networks — limited-attention Alpha-Centrality ($laAC$) — that model attention-limited nature of social interactions and provide their mathematical definitions. We also developed fast approximate algorithm to calculate this measure on large graphs and provided its performance guarantees.

Alpha-Centrality measures the total number of paths from a node, exponentially attenuated by their length. Bonacich introduced this measure as a generalization of the index of status proposed by Katz, and it is sometimes referred to as Bonacich centrality. Alpha-Centrality vector $\mathrm{cr}(\alpha, s)$ can be defined iteratively in terms of adjacency matrix of the graph A:

$$\mathrm{cr}(\alpha, s) = s + \alpha A \cdot \mathrm{cr}(\alpha, s), \tag{3}$$

where the starting vector $s = Ae^T$ is taken as out-degree centrality [2].

Alpha-Centrality gives the steady state distribution of an epidemic process on a network [10], where α is the probability to transmit a message or influence along a link. Therefore, ith entry of cr can be interpreted as the number of infections directly or indirectly caused by node i (see attached paper for more details).



(a) AC (b) laAC

Figure 4: Directed network with sizes of nodes weighed by their score according to (a) Alpha-centrality and (b) limited-attention Alpha-centrality of the influence graph.

Let us now consider the case in which a node's capacity to receive incoming stimuli — whether messages or viruses — is limited. While attention need not be distributed uniformly over friends — some friends may receive a greater share of a person's attention due to familiarity, trust, social closeness, or influence — for simplicity, we assume that each friend receives the same fraction of a person's attention. Therefore, the probability that node j will receive a message broadcast by i will be proportional to $1/d_{in}(j)$, where $d_{in}(j)$ is the in-degree of node j. The limited-attention Alpha-Centrality matrix can be written in terms of the modified adjacency matrix $M = AD_{in}^{-1}$ as:

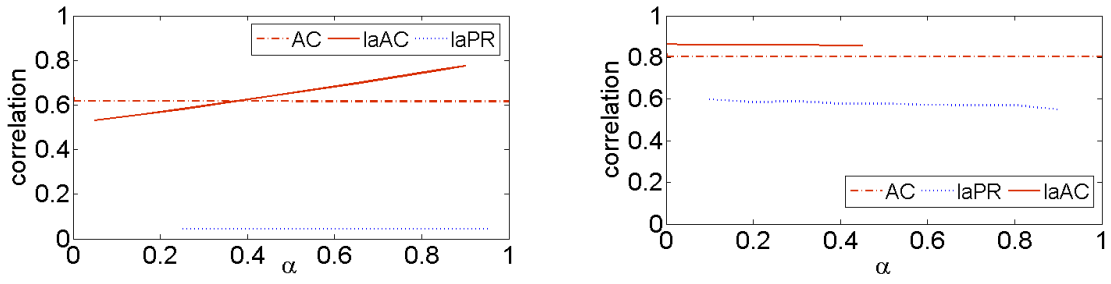
$$C_{la} = M + \alpha M^2 + \alpha^2 M^3 + \alpha^3 M^4 + \dots$$

The limited-attention Alpha-Centrality vector ${}^{la}\mathrm{cr}(\alpha, s)$ can also be written in iterative form:

$${}^{la}\mathrm{cr}(\alpha, s) = s + \alpha A D_{in}^{-1} \cdot {}^{la}\,\mathrm{cr}(\alpha, s), \tag{4}$$

with the starting vector $s = AD_{in}^{-1}e^T$. Figures 4 illustrates the differences between Alpha-Centrality and its limited-attention variant.

Note that we have developed fast approximate algorithms to compute Alpha-Centrality and limited-attention Alpha-Centrality.



(a) Digg (b) Twitter

Figure 5: Correlation of rankings of (a) Digg and (b) Twitter users found by different measures of centrality with the empirical influence ranking.

We evaluated the performance of the centrality measures on the task of identifying influential users of large networks from social media sites containing hundreds of thousands of users. However, in order to evaluate the performance of centrality, we need a relevant measure of influence. User activity provides us with an empirical measure of influence. When a user posts a URL on Digg or Twitter, she broadcasts it to all her followers. We refer to this user as the *submitter*. Whether or not her follower will re-broadcast the URL (i.e., retweet it on Twitter or vote for it on Digg) depends on its *quality* and *submitter's influence*. Assuming that URL's quality is uncorrelated with the submitter, we can average out its effect by aggregating over all URLs submitted by the same user. The residual difference in the amount of attention the followers pay submitter by re-broadcasting her messages can be attributed to variations in submitter's influence. Therefore, we use the average number of times the URLs submitted by the user are re-broadcast by her followers as the *empirical measure of influence*.

Figure 5 shows how well the rankings produced by different centralities correlate with the empirical influence rankings of users who submitted at least two URLs which were rebroadcast at least ten times. We use Spearman rank correlation because it is less sensitive to variations in scores, and we expect some variation to arise in approximate centrality scores. Limited-attention Alpha-Centrality correlates better with the empirical measure of influence than Alpha-Centrality over a broad range of α values, consistent with our claim that $laAC$ is a better measure for predicting central social media users, because it better models the dynamics of online communication than AC. On Digg, AC appears to outperform $laAC$ for small values of α. Since α can be thought of as the scale of interaction, this implies that locally, AC better predicts influential users. This could be the consequence of the fact that our measure of influence, i.e., number of re-broadcasts by followers, is a local measure. In the future, we plan to compare the performance of centrality measures using a global measure of influence, for example, the average size of cascades triggered by submitted URLs. We did not expect limited-attention PageRank ($laPR$) (described in the accompanying paper) to predict influence rankings of Digg and Twitter users, since the dynamic process this centrality models does not at all describe communication patterns of social media users, and we found no correlation.

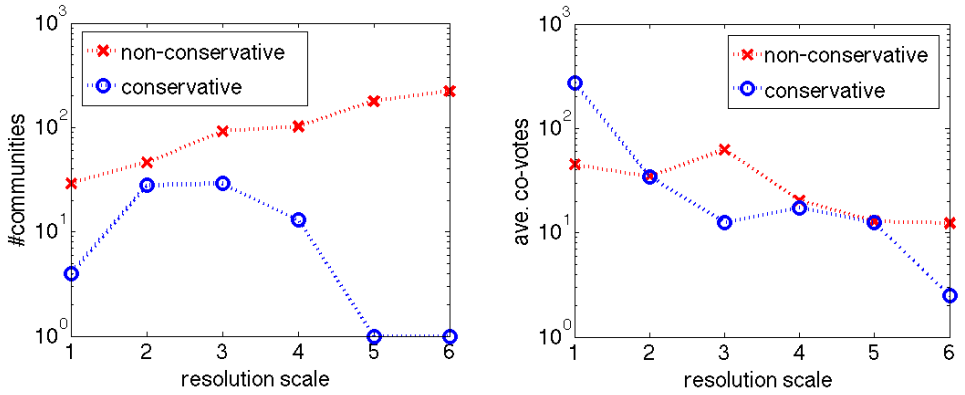
3.1.1 Relevant Publications.

- R. Ghosh and K. Lerman. Rethinking centrality: The role of dynamical processes in social network analysis. *Discrete and Continuous Dynamical Systems Series B*, 19(5):1355 – 1372, July 2014.
- Lerman, K.; Jain, P.; Ghosh, R.; Kang, J.; and Kumaraguru, P. Limited Attention and Centrality in Social Networks In *Proceedings of International Conference on Social Intelligence and Technology* (SOCIETY), 2013.
- R. Ghosh and K. Lerman. Parameterized centrality metric for network analysis. *Physical Review E*, 83(6):066118, June 2011.

3.2 Dynamics and Communities

Dynamics processes also affect the emergent communities. We illustrate on a simple example of opinion formation. Imagine a network of interacting agents, each holding an opinion. Each interaction causes agents' opinions to become more similar. As the network evolves over time, opinions of agents within the same community converge faster than those of other agents. This framework allows us to study how network topology and interactions, which mediate the transfer of opinions between agents, both affect the formation of communities. In traditional models of opinion dynamics, agents interact via one-to-one opinion transfer. Such conservative interactions can be modeled as random walks. However, social interactions are often non-conservative, resulting in one-to-many transfer of opinions. These interactions result in different emergent patterns of consensus — or communities of agents holding the same opinion.

We simulated dynamics of opinion formation in the real-world networks of the Digg follower graph and Facebook friendship graph. We simulated two different interaction models: one-to-one and one-to-many interactions. Both types of models revealed similar "core and whiskers" community structure in each network, with a giant core and small communities, or whiskers, loosely attached to the core. Furthermore, this structure was multi-scale. Isolating the core, and simulating dynamics of opinion formation within it revealed finer-grained structure, with another core and small "whiskers" attached to it. However, the composition of the giant cores identified by the two interaction models were very different.



(a) (b)

Figure 6: Properties of small communities found in the Digg mutual follower graph by the two interaction models. (a) Number of small communities at different resolutions scales. (b) Average number of co-votes made by community members.

Figure 6(a) shows the number of small communities (whiskers) resolved by the two interaction models at different resolution scales, measured by closeness to the center of the core. One-to-many (non-conservative) interaction model assigned many more users to such communities than the one-to-one (conservative) interaction model. The rest of the users fragmented into isolated pairs or singletons, who did not synchronize their opinions with any others. How does the *quality* of the discovered communities differ? We measure similarity of two Digg users by the number of stories for which they both voted, i.e., *co-votes*. Then, averaging over co-votes of all connected pairs of community members, we obtain a measure of community "cohesiveness." As seen in Figure 6(b), average number of co-votes increases at finer resolution scales, producing more cohesive communities in the center of the core.

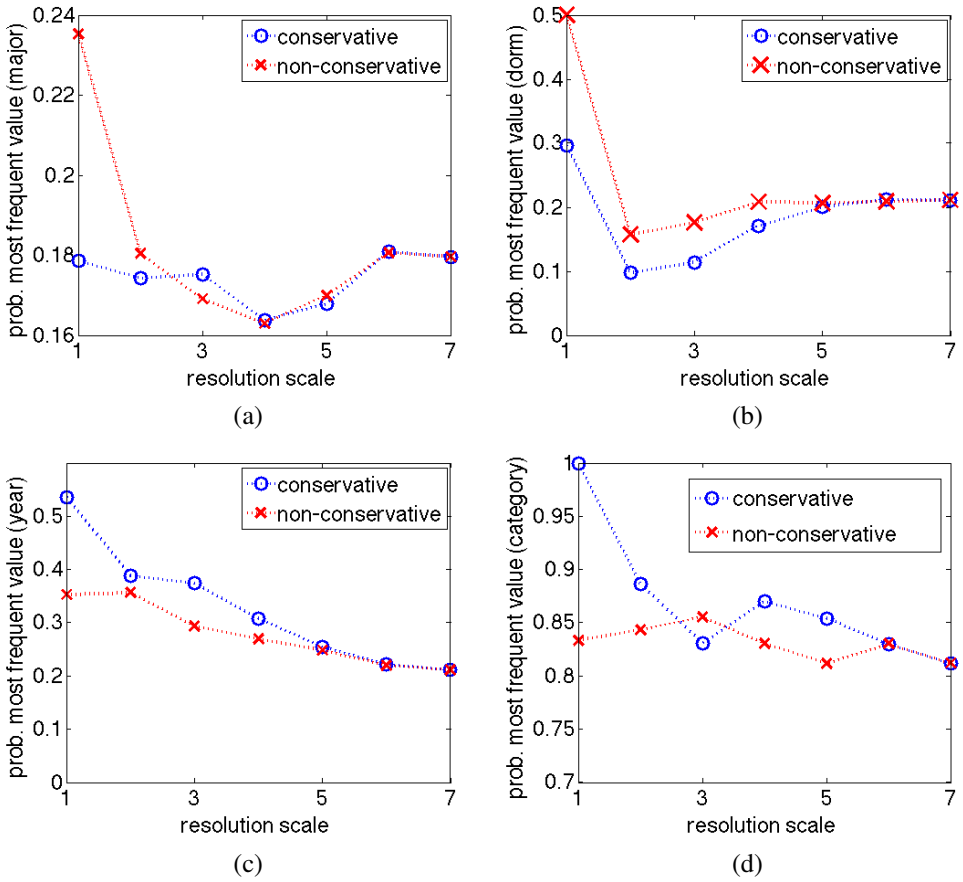


Figure 7: Properties of small communities found in the Facebook network of American University by the two interaction models. Each plot shows at different resolution scales the probability of occurrence of the most frequent value of user features (a) major, (b) dorm, (c) year, (d) category of individual. Conservative and non-conservative refer to one-to-one and one-to-many interactions respectively.

How do the small communities discovered at different resolution scales by the two interaction models in the Facebook social network differ? We look at four features of users in the data set — major, dorm, year and category of individual — and calculate the prevalence of feature values among community members. The community is characterized by the prevalence of the most popular feature among its members, or its cohesiveness with respect to that feature. For example, when using the dorm feature to characterize the community, dorm cohesiveness is the largest fraction of community members that belong to the same dorm. Figure 7 shows the cohesiveness of communities found by the two interaction models at different resolution scales with respect to some feature generally increase at finer resolution scale. This suggests that individuals in the center of the core are far more similar to each other than peripherally connected individuals. More importantly, the characteristics of the community structure discovered by conservative and non-conservative interaction models vary significantly.

3.2.1 Relevant Publications.

- R. Ghosh and K. Lerman. The Impact of Network Flows on Community Formation in Models of Opinion Dynamics. to appear in *J. of Mathematical Sociology*.
- L. M. Smith, K. Lerman, C. Garcia-Cardona, A. G. Percus, and R. Ghosh. Spectral clustering with epidemic diffusion. *Physical Review E*, 88(4):042813, Oct. 2013.
- K. Lerman and R. Ghosh. Network Structure, Topology and Dynamics in Generalized Models of Synchronization. *Physical Review E*, 86(026108), 2012.

3.3 Dynamics and Link Prediction

As described above, predicting new links in networks is a critical capability, both for the military and for commercial applications, such as product and social recommendation. Link prediction heuristics take network's current structure into account to predict new links that will form between existing nodes. A variety of link prediction heuristics have been proposed, including neighborhood overlap, the Adamic-Adar score, which weighs the contribution of each common neighbor by the inverse of the logarithm of its degree, and number of paths connecting the two nodes. In general, link prediction heuristics consider how close two nodes are in a network. In the money exchange network, for example, two individuals can be considered to be close even if they don't know each other, if the money distributed by one is often received by the other. In other words, the probability that a random walk originating at one node reaches the other measures the proximity of two nodes in the network.

We argue that proximity should take into account the ability to exchange information or to influence with each other. This is determined by the dynamic processes taking place on the network, i.e., the processes by which information or influence is transmitted from one node to another.

In this project, we unified link prediction heuristics by viewing them as instances of network proximity under different dynamics, and introduced new ones based on other dynamic processes. The heuristics we examined are listed in Table 2. Note, that these are defined for directed networks, where $\Gamma_{in}(v)$ refers to the in-neighbors of node v, i.e., $d_{in}(v)$ nodes whose edges are incident on v, where $d_{in}(v)$ is the in-degree of v, and similarly for the out-neighbors. In addition to existing ones, we introduced new heuristics based on node's limited bandwidth. Consider a process in which a node's capacity to receive incoming messages is limited by its bandwidth. As a consequence, the more incoming connections (in-links) a node has, the less likely it is to receive a message from an arbitrary connection, e.g., because it has already reached the limit of its capacity by processing other incoming messages. This alters the character of the flow and leads to novel measures of network proximity, that we call limited-bandwidth or limited-attention epidemics or random walks.

Table 2: Heuristics used in link prediction applications. Popular existing link prediction heuristics appear above the double line: number of common neighbors, Jaccard and Adamic-Adar score, and resource allocation. Below the double line are link prediction heuristics introduced in this paper.

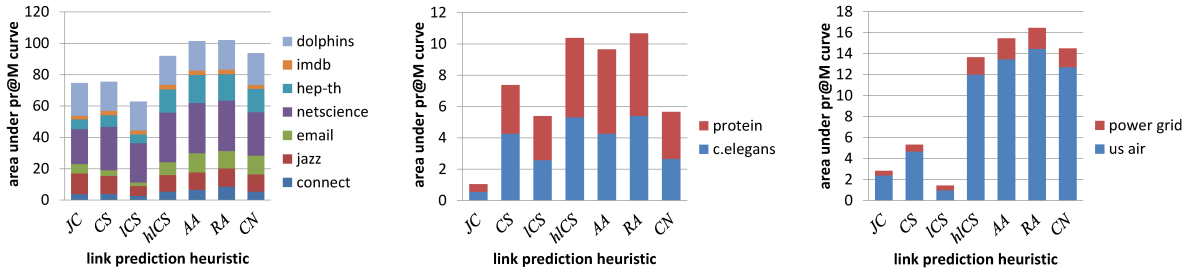
name	**symbol**	**definition**
common neighbors	CN	$CN = \frac{1}{2}\left[\lvert\Delta\rvert + \lvert\Delta'\rvert\right]$
Jaccard score	JC	$JC = \frac{1}{2}\left[\frac{\lvert\Gamma_{\rm out}(u)\cap\Gamma_{\rm in}(v)\rvert}{\lvert\Gamma_{\rm out}(u)\cup\Gamma_{\rm in}(v)\rvert} + \frac{\lvert\Gamma_{\rm out}(v)\cap\Gamma_{\rm in}(u)\rvert}{\lvert\Gamma_{\rm out}(v)\cup\Gamma_{\rm in}(u)\rvert}\right]$
Adamic-Adar	AA	$AA = \frac{1}{2}\left[\sum_{z\in\Delta}\frac{1}{\log(d(z))} + \sum_{z'\in\Delta'}\frac{1}{\log(d(z'))}\right]$
resource allocation	RA	$RA = \frac{1}{2}\left[\sum_{z\in\Delta}\frac{1}{d(z)} + \sum_{z'\in\Delta'}\frac{1}{d(z')}\right]$
conservative (random walk)	CS	$CS = \frac{1}{2}\sum_{z\in\Delta}\frac{1}{d_{\rm out}(u)d_{\rm out}(z)} + \frac{1}{2}\sum_{z\in\Delta'}\frac{1}{d_{\rm out}(v)d_{\rm out}(z)}$
limited-bandwidth conservative	lCS	$lCS = \frac{1}{2}\sum_{z\in\Delta}\frac{1}{d_{\rm out}(u)d_{\rm in}(z)d_{\rm out}(z)d_{\rm in}(v)} + \frac{1}{2}\sum_{z\in\Delta'}\frac{1}{d_{\rm out}(v)d_{\rm in}(z)d_{\rm out}(z)d_{\rm in}(u)}$
non-conservative (epidemic)	NC	$NC = \frac{1}{2}\left[\lvert\Delta\rvert + \lvert\Delta'\rvert\right]$
limited-bandwidth non-conservative	lNC	$lNC = \frac{1}{2}\sum_{z\in\Delta}\frac{1}{d_{\rm in}(z)d_{\rm in}(v)} + \frac{1}{2}\sum_{z\in\Delta'}\frac{1}{d_{\rm in}(z)d_{\rm in}(u)}$
hybrid conservative	hCS	$hCS = \frac{1}{2}\left[\sum_{z\in\Delta}\frac{1}{d_{\rm out}(z)} + \sum_{z\in\Delta'}\frac{1}{d_{\rm out}(z)}\right]$
hybrid limited-bandwidth conservative	$hlCS$	
hybrid non-conservative	lNC	
hybrid limited-bandwidth non-conservative	$hlNC$	

Table 3: Networks studied in the missing link prediction task and their properties.

network	nodes	edges	missing	density
social networks				
dolphins	62	159	16	0.084
email	1133	5452	545	0.0085
jazz	198	2742	274	0.14
connect	1095	7825	783	0.014
hep-th	8710	14254	1425	0.0003
netscience	1461	2742	274	0.0013
imdb	6260	98235	9824	0.005
technological networks				
us air	332	2126	212	0.0193
power grid	4941	6594	660	0.0004
biological networks				
protein	1870	2277	228	0.0013
c. elegans	453	2040	204	0.02

We conducted experiments on a variety of networks belonging to three categories: $Social$, $Technological$ and $Biological$ networks. Table 3 lists some of the statistics of the datasets. We evaluated the performance of link prediction heuristics on the missing link prediction task in these networks. Since all the networks studied here are undirected, some of the heuristics are mathematically equivalent: $CS = lNC$, $RA = hCS = hlNC$, and $CN = NC = hNC$.

We ran several trials of the link prediction task for each network. In each trial, first, we randomly remove 10% of all edges and assign them to the test set E_{test}. The remaining 90% of links comprise the training set, the graph $G_{train} = (V, E_{train})$. We then compute network proximity using a given link prediction heuristic for all pairs of nodes $|V \times V - E_{train}|$ and rank them in decreasing order. We score the prediction based on how many of top-M predicted edges are correct. This allows us to compute a curve showing *precision@M*, the ratio of the number of correctly predicted links within the M edges with the highest score.



social networks biological networks technological networks

Figure 8: Aggregated performance of different link prediction heuristics.

We compare the performance of different link prediction heuristics by measuring the area under the precision curve. Figure 8 aggregates these measure across all datasets within each domain, giving us a sense of their relative performance. First thing we note is that there is a wide variation in performance of different link prediction metrics, we a popular Jaccard similarity performing poorly in many networks. However, some metrics perform consistently better: RA, AA and $hlCS$ are consistently among the top performing measures. RA measure is related to random-walk based measure CS and limited-attention epidemic, and both AA and $hlCS$ are variants of this measure. Though we cannot confirm whether it is the random walk or the limited attention that leads to better performance (since random walk is mathematically equivalent to a limited-attention epidemic in an undirected graph), this study indicates that a practioner should be careful about choosing an appropriate metric for the task, one that reflects the phenomena taking place in the network.

3.3.1 Relevant Publications.

- Narang, K.; Lerman, K.; and Kumaraguru, P. Network Flows and the Link Prediction Problem In *Proceedings of KDD workshop on Social Network Analysis* (SNA-KDD), 2013.

3.4 Generalized Laplacian Framework

Our major accomplishment was the mathematical framework to model dynamics on networks. This framework unifies several well known centrality and community measures under a single model. In this dynamics-oriented view, a node's centrality describes its participation in the dynamical process taking place on the network. Similarly, communities are groups of nodes that interact more frequently with each other [9].

The mathematical framework we developed is general and flexible, able to represent a variety of dynamical processes. At the core of this framework is the *generalized Laplacian* matrix:

$$\boldsymbol{\mathcal{L}} = (\boldsymbol{T}\boldsymbol{D}_{\boldsymbol{W}})^{-1/2-\rho}(\boldsymbol{D}_{\boldsymbol{W}} - \boldsymbol{W})(\boldsymbol{D}_{\boldsymbol{W}}\boldsymbol{T})^{-1/2+\rho}. \tag{5}$$

Compared with the traditional symmetric normalized Laplacian $\boldsymbol{D}^{-1/2}(\boldsymbol{D} - \boldsymbol{A})\boldsymbol{D}^{-1/2}$, generalized Laplacian has three additional parameters corresponding to different linear transformations. These transformations relate the different dynamical processes to the random walk.

1. ρ: Similarity transformation. It is an equivalence relation on the space of square matrices, leading to seemingly unrelated formulations which are in fact the same dynamics under different basis. While ρ technically can be any real number, in this work we limit ourselves to three special cases: $\rho = 1/2, 0, -1/2$. These cases correspond to three equivalent formulations we shall call "consensus"($\boldsymbol{\mathcal{L}}^{CON}$), "symmetric"($\boldsymbol{\mathcal{L}}^{SYM}$) and "random walk"($\boldsymbol{\mathcal{L}}^{RW}$) respectively. While we use $\boldsymbol{\mathcal{L}}^{SYM}$ for mathematical convenience, it is often more intuitive to think from the random walk or consensus perspective.

2. $\boldsymbol{T}$: Scaling transformation. $\boldsymbol{T}$ is the $n \times n$ diagonal matrix of *vertex delay factors*. Its ith element τ_i represents the average delay of vertex i. Without loss of generality, we assume that $\tau_i \geq 1$, for all $i \in V$. Scaling transformation can be understood as rescaling the local delay at each vertex i by τ_i, with the dynamical process's waiting times between jumps exponentially distributed as the PDF $f(t, \tau) = \frac{1}{\tau_i} e^{-\frac{t}{\tau_i}}$.

3. $\boldsymbol{W}$: Reweighing transformation that gives new weights to edges of the network. We use the *interaction matrix* $\boldsymbol{W}$ instead of the adjacency matrix $\boldsymbol{A}$. Note that the degree matrix $\boldsymbol{D}_{\boldsymbol{W}}$ is now also defined in terms of $\boldsymbol{W}$, that is $d_{\boldsymbol{W}\,i} = \sum_j w_{i,j}$. While the scaling transformation changes the delay at each vertex, reweighing transformation changes the trajectory of a dynamic process. For example, a biased random walk with transition probability $P_{ij} \propto \alpha_i \boldsymbol{A}_{ij}$ is equivalent to an unbiased random walk on the reweighed "interaction graph" $\boldsymbol{W}$ with entries $w_{ij} = \alpha_i \boldsymbol{A}_{ij} \alpha_j$.

The generalized Laplacian framework is flexible enough to capture a variety of well-known processes, such as random walks and epidemics, but also describe less-studied processes.

Normalized Laplacian If the interaction matrix is the adjacency matrix $\boldsymbol{W} = \boldsymbol{A}$ and vertex delay factor is the identity $\boldsymbol{T} = \boldsymbol{I}$, with $\rho = 0$ we recover the *symmetric normalized Laplacian*:

$$\boldsymbol{\mathcal{L}}^{SYM} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}.$$

Under similarity transformations, normalized Laplacian is equivalent to some well studied dynamical processes. For example, by setting $\rho = -1/2$ we have the *unbiased random walk*

$$\boldsymbol{\mathcal{L}}^{RW} = \boldsymbol{I} - \boldsymbol{A}\boldsymbol{D}^{-1} \,,$$

where $\boldsymbol{A}\boldsymbol{D}^{-1}$ forms a stochastic matrix whose $ijth$ entry is the transition probability P_{ij}. Setting $\rho = 1/2$ we have the *consensus* formulation

$$\boldsymbol{\mathcal{L}}^{CON} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A} \,,$$

where each vertex updates its "belief" based on the weighted average "beliefs" of its neighbors.

(Scaled) Graph Laplacian When $\boldsymbol{W} = \boldsymbol{A}$, $\boldsymbol{T} = d_{max}\boldsymbol{D}^{-1}$, the generalized Laplacian operator corresponds to the (scaled) graph *Laplacian*

$$\boldsymbol{\mathcal{L}} = 1/d_{max}(\boldsymbol{D} - \boldsymbol{A}).$$

Notice that by setting $\boldsymbol{T} = d_{max}\boldsymbol{D}^{-1}$, the diagonal matrix $\boldsymbol{T}\boldsymbol{D}_{\boldsymbol{W}}$ becomes effectively a scalar. As a result, different similarity transformation (changing ρ) lead to identical linear operators. This operator is often used to describe *heat diffusion* processes and its $\boldsymbol{\mathcal{L}}^{CON}$ interpretation is used for distributed calculation of arithmetic means.

Replicator Let $\boldsymbol{v}$ be the eigenvector of $\boldsymbol{A}$ associated with its largest eigenvalue λ_{max}: $\boldsymbol{A}\boldsymbol{v} = \lambda_{max}\boldsymbol{v}$. We can then construct a diagonal matrix $\boldsymbol{V}$ whose elements are the components of the eigenvector $\boldsymbol{v}$. Let us consider the interaction matrix $\boldsymbol{W} = \boldsymbol{V}\boldsymbol{A}\boldsymbol{V}$ with $\boldsymbol{T} = \boldsymbol{I}$ and $\rho = 0$:

$$\boldsymbol{\mathcal{L}}^{SYM} = \boldsymbol{I} - \boldsymbol{D}_{\boldsymbol{W}}^{-1/2}\boldsymbol{W}\boldsymbol{D}_{\boldsymbol{W}}^{-1/2} = \boldsymbol{I} - 1/\lambda_{max}\boldsymbol{A}$$

This operator is known as the replicator matrix $\boldsymbol{R}$, and it models *epidemic diffusion* on a graph [17, 18]. Setting $\rho = -1/2$ we get the maximum entropy random walk on the original graph $\boldsymbol{A}$.

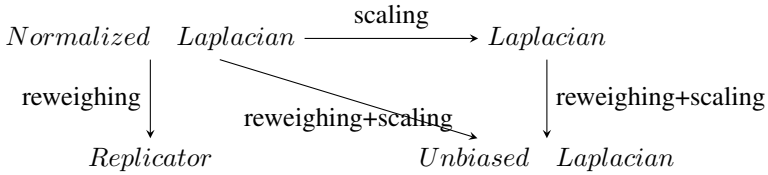
$$\boldsymbol{\mathcal{L}}^{RW} = \boldsymbol{I} - \boldsymbol{W}\boldsymbol{D}_{\boldsymbol{W}}^{-1} \ .$$

Unbiased Laplacian Normalized adjacency matrix is known as $\boldsymbol{W} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$. With $\boldsymbol{T} = d_{\boldsymbol{W}\,max}D_{\boldsymbol{W}}^{-1}$ we define the *unbiased Laplacian matrix*:

$$\boldsymbol{\mathcal{L}} = 1/d_{\boldsymbol{W}\,max}(\boldsymbol{D}_{\boldsymbol{W}} - \boldsymbol{W}).$$

Just like the graph Laplacian, different values of ρ for Unbiased Laplacian lead to the same operator. Its $\boldsymbol{\mathcal{L}}^{RW}$ interpretation is a degree based biased random walk with $P_{ij} \propto d_i^{-1/2}\boldsymbol{A}_{ij}$.

These four special cases are related to each other through scaling and reweighing transformations, captured by the following diagram.



Our empirical study demonstrates that these special cases and different transformations in general can lead to divergent views about who the central vertices are and what are the corresponding communities (see figures). The above diagram helps us better understand how different measures of centrality and communities relate to each other under the generalized Laplacian framework.

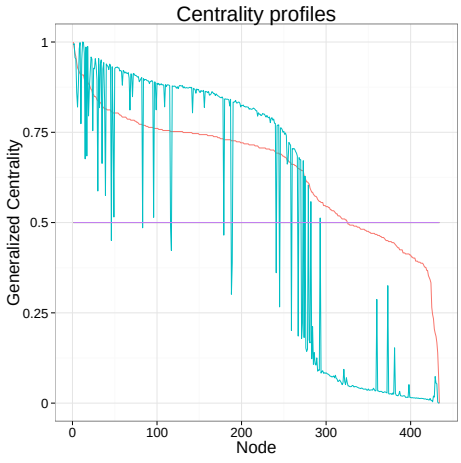
3.4.1 Generalized Centrality.

Centrality captures how important a node is in a network. Under the generalized Laplacian framework, different centrality measures are related to solutions of different dynamical processes [10].

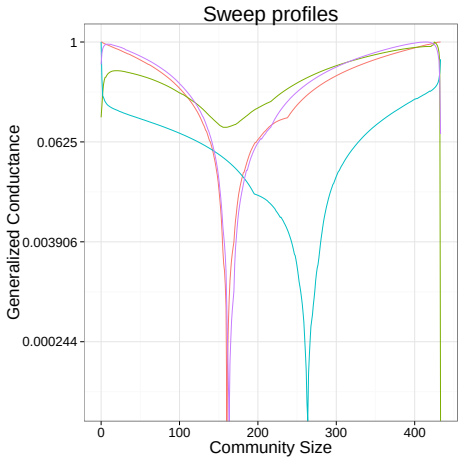
Similarity transformations (different values of ρ) lead to the same state vector $\boldsymbol{\theta}$ at any time t up to a change of basis. Based on the connection between centrality measures and the stationary distribution of a random walk, we generalize the definition of centrality to:

$$c_i = d_{\boldsymbol{W}\,i}\tau_i \ . \tag{6}$$

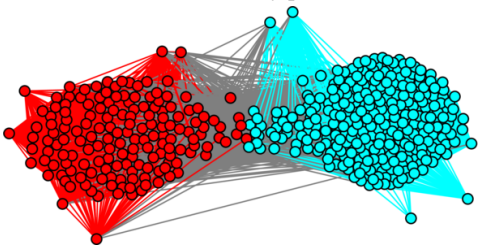
Generalized centrality reduces to some well known centrality measures by setting the parameters $\boldsymbol{T}$ and $\boldsymbol{W}$. They can now be systematically compared by scaling and reweighing transformations between special cases under the generalized Laplacian framework. They include *degree centrality* d_i for Normalized Laplacian, and the square of *eigenvector centrality* $\boldsymbol{v}_i^2$ for Replicator. First column of the figures illustrate how generalized centrality differ in four special cases on the same network.



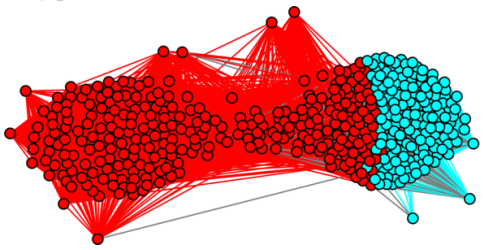
(a) Centrality profile



(b) Community profile



(c) Spectral partition using Normalized Laplacian



(d) Spectral partition using Replicator

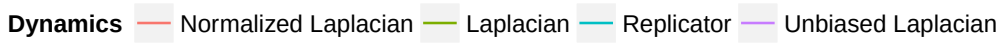
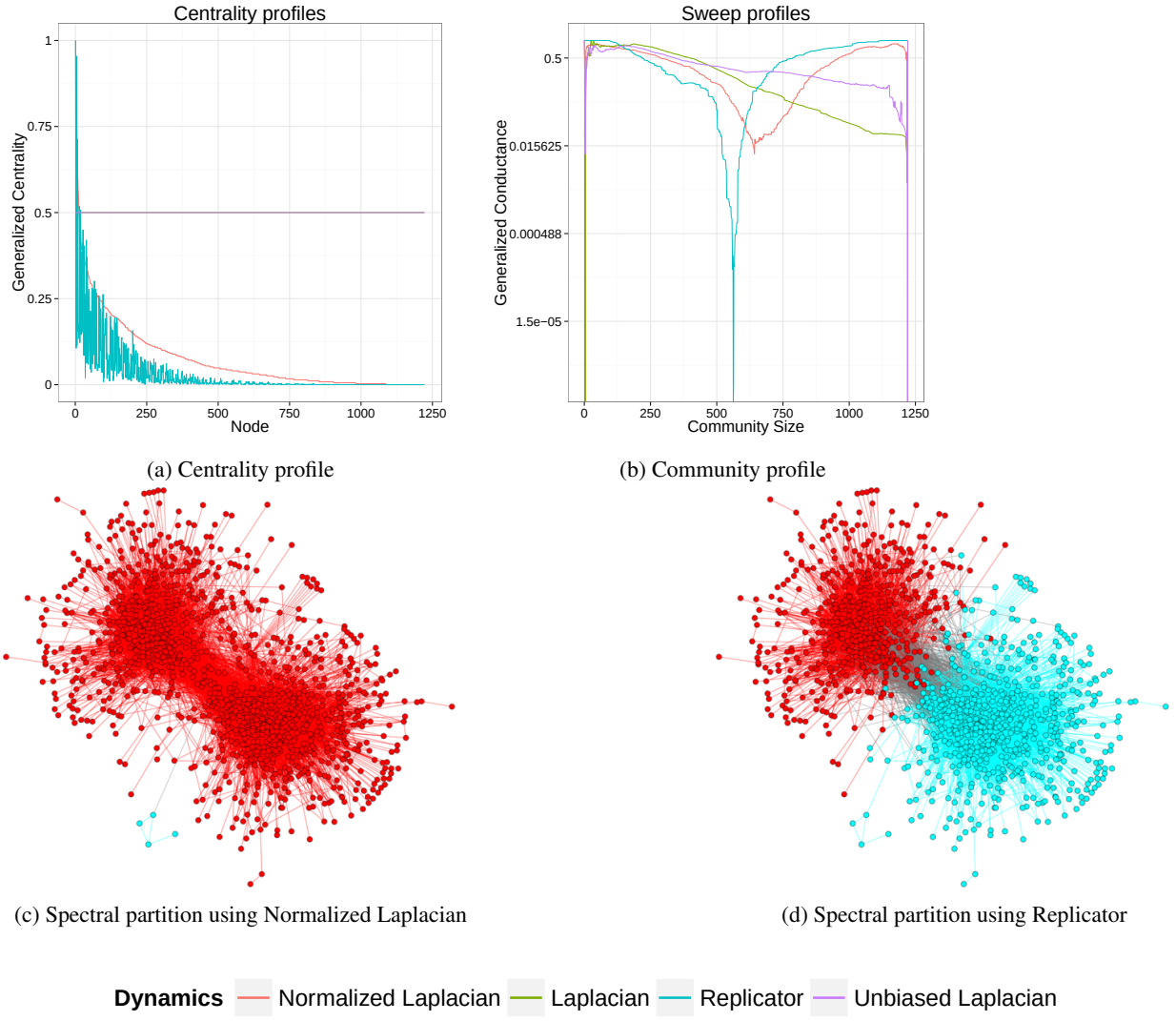


Figure 9: Different views of the structure of the House of Representatives network (centrality and communities) resulting from the dynamics specified by different operators.



(a) Centrality profile

(b) Community profile

(c) Spectral partition using Normalized Laplacian

(d) Spectral partition using Replicator

Figure 10: Different views of the structure of the Political Blogs network (centrality and communities) resulting from the dynamics specified by different operators.

3.4.2 Generalized Conductance and Spectral Clustering.

With generalized Laplacian framework, we can also define a *generalized conductance* that measures the quality of a subset S (of vertices) as a potential community. This measure is used in spectral clustering to find optimal cut of the network into subgraphs representing different communities.

$$h_{\boldsymbol{\mathcal{L}}}(S) = \frac{\mathrm{cut}(S, \bar{S})}{\min\left(\mathrm{vol}_{\boldsymbol{\mathcal{L}}}(S), \mathrm{vol}_{\boldsymbol{\mathcal{L}}}(\bar{S})\right)} = \frac{\sum_{i \in S, j \in \bar{S}} w_{i,j}}{\min\left(\sum_{i \in \mathrm{S}} d_{\boldsymbol{W}\, i} \tau_i, \sum_{i \in \bar{\mathrm{S}}} d_{\boldsymbol{W}\, i} \tau_i\right)} \,. \tag{7}$$

Notice that we have generalized the volume measure of a set $S \subseteq V$ to $\mathrm{vol}_{\boldsymbol{\mathcal{L}}}(S) = \sum_{i \in S} d_{\boldsymbol{W}\, i} \tau_i$, which is the sum of generalized centrality of member vertices.

With generalized conductance we can extend the classic Cheeger's inequality, which relates the second smallest eigenvalue of the normalized Laplacian to the conductance of the best bisection in the network, to their generalized counterparts under our framework.

$$(\min_{S \in V} h_{\boldsymbol{\mathcal{L}}}(S))^2/2 \leq \lambda_1 \leq 2 \min_{S \in V} h_{\boldsymbol{\mathcal{L}}}(S) \tag{8}$$

These theoretical results eventually lead to efficient spectral clustering algorithm for detecting communities associated with different dynamics. Our framework also paves the way for efficient local graph partitioning. Last three columns of the figures illustrate how generalized conductance of the good partitions found by our algorithm differ in four special cases on the same network.

Figure 9 and 10 show the centrality profile and community profile obtained by different operators on two benchmark networks. The first network, House of Representatives, shows the network of congressmen, where a link represents a co-vote on a bill. The second network, political blogs, shows hyperlinks between blogs. Centrality profile gives generalized centrality for each node given a dynamic operator. The community sweep profile in the second column gives generalized conductance, for a potential community of size k using our spectral clustering algorithm [9]. To improve visualization, both are reordered on the x axis and rescaled on the y axis. The visualizations in the last two columns are the best bisections found by our algorithm using the indicated special case. As can be seen from the figures, the centrality profiles under different operators are very different, resulting in alternate visions of who the important nodes are. In addition, community sweep profiles are also very different, and lead to different partitions of the same network into communities.

3.4.3 Relevant Publications.

- R. Ghosh, K. Lerman, S.-H. Teng, and X. Yan. The interplay between dynamics and networks: Centrality, communities, and cheeger inequality. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2014)*, 2014.

4 DELIVERABLES

Over the course of the project, we delivered code to AFRL. The first deliverable was Netkit, a network analysis toolbox that includes several centrality computation methods.

The second deliverable was link prediction code that takes as input a network and returns a ranked list of links more likely, but not yet observed, links.

5 PUBLICATIONS

The work conducted over the course of this project resulted in seven journal publications and many more papers in conference proceedings.

References

[1] L. Katz. A new status derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

[2] Phillip Bonacich. Power and centrality: A family of measures. *The American Journal of Sociology*, 92(5):1170–1182, 1987.

[3] P.B. Bonacich. Eigenvector-like measures of centrality for assymetric relations. *Social Networks*, 23:191–201, 2001.

[4] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[6] Rumi Ghosh and Kristina Lerman. Parameterized centrality metric for network analysis. *Physical Review E*, 83(6):066118, June 2011.

[7] Rumi Ghosh, Tawan Surachawala, and Kristina Lerman. Entropy-based classification of retweeting activity on twitter. In *Proceedings of KDD workshop on Social Network Analysis (SNA-KDD)*, August 2011.

[8] Rumi Ghosh and Kristina Lerman. Predicting influential users in online social networks. In *Proceedings of KDD workshop on Social Network Analysis (SNA-KDD)*, July 2010.

[9] Rumi Ghosh, Kristina Lerman, Shang-Hua Teng, and Xiaoran Yan. The interplay between dynamics and networks: Centrality, communities, and cheeger inequality. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2014)*, 2014.

[10] Rumi Ghosh and Kristina Lerman. Rethinking centrality: The role of dynamical processes in social network analysis. *Discrete and Continuous Dynamical Systems Series B*, 19(5):1355 – 1372, July 2014.

[11] Kristina Lerman, Rumi Ghosh, and Jeon-hyung Kang. Centrality metric for dynamic network analysis. In *Proceedings of KDD workshop on Mining and Learning with Graphs (MLG)*, July 2010.

[12] Rumi Ghosh, Tsung-Ting Kuo, Chun-Nan Hsu, Shou-De Lin, and Kristina Lerman. Time-aware ranking in dynamic citation networks. In *COMMPER 2011: Mining Communities and People Recommendations, Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 373 –380, December 2011.

[13] Yi hung Huang, Chun-Nan Hsu, and Kristina Lerman. Identifying transformative scientific research. In *Proc. of IEEE International Conference on Data Mining*, 2013.

[14] Rumi Ghosh and Kristina Lerman. A framework for quantitative analysis of cascades on networks. In *Proceedings of Web Search and Data Mining Conference (WSDM)*, February 2011.

[15] Linhong Zhu and Kristina Lerman. A visibility-based model for link prediction in social media. In *Proceedings of the ASE/IEEE Conference on Social Computing*, 2014.

[16] Nathan O. Hodas and Kristina Lerman. How limited visibility and divided attention constrain social contagion. In *ASE/IEEE International Conference on Social Computing*, 2012.

[17] Kristina Lerman and Rumi Ghosh. Network Structure, Topology and Dynamics in Generalized Models of Synchronization. *Physical Review E*, 86(026108), 2012.

[18] L. M. Smith, K. Lerman, C. Garcia-Cardona, A. G. Percus, and R. Ghosh. Spectral clustering with epidemic diffusion. *Physical Review E*, 88(4):042813, October 2013.

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

$\boldsymbol{A}$	adjacency matrix of a network
$\boldsymbol{D}$	diagonal out-degree matrix
$\boldsymbol{D}_W$	out-degree matrix of the reweighted network
$\boldsymbol{T}$	node scaling (time delay) factor matrix
$\boldsymbol{I}$	identify matrix
C	centrality matrix
cr	centrality vector
α	parameter in centrality calculations setting the length scale of interactions
$laAC$	limited-attention Alpha-Centrality
CN	link prediction heuristic: number of common neighbors
JC	link prediction heuristic: Jaccard score
AA	link prediction heuristic: Adamic-Adar score
CS	link prediction heuristic: conservative score (random walk)
lCS	link prediction heuristic: limited-bandwidth conservative score
NC	link prediction heuristic: non-conservative score (epidemic)
lNC	link prediction heuristic: limited-bandwidth non-conservative
hCS	link prediction heuristic: hybrid conservative
$hlCS$	link prediction heuristic: hybrid conservative
hNC	link prediction heuristic: hybrid non-conservative
$hlNC$	link prediction heuristic: hybrid limited-bandwidth non-conservative
$\boldsymbol{\mathcal{L}}$	Laplacian matrix of the network specifying dynamics
$\boldsymbol{\mathcal{L}}^{SYM}$	normalized symmetric Laplacian
$\boldsymbol{\mathcal{L}}^{RW}$	random-walk Laplacian
$\boldsymbol{\mathcal{L}}^{CON}$	consensus Laplacian
$h_{\boldsymbol{\mathcal{L}}}$	conductance of the network with respect to dynamics $\boldsymbol{\mathcal{L}}$
$\mathrm{vol}_{\boldsymbol{\mathcal{L}}}$	volume of the network with respect to dynamics $\boldsymbol{\mathcal{L}}$