The Role of Information Diffusion in the Evolution of Social Networks

Lilian Weng[1], Jacob Ratkiewicz[2], Nicola Perra[3], Bruno Gonçalves[4], Carlos Castillo[5], Francesco Bonchi[6], Rossano Schifanella[7], Filippo Menczer[1], Alessandro Flammini[1]
[1]School of Informatics and Computing, Indiana University Bloomington, USA [2]Google Inc.
[3]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, USA
[4]Aix Marseille Université, CNRS, CPT, UMR 7332, Marseille, France
[5]Qatar Computing Research Institute [6]Yahoo! Research Barcelona
[7]Department of Computer Science, University of Torino, Italy

ABSTRACT

Every day millions of users are connected through online social networks, generating a rich trove of data that allows us to study the mechanisms behind human interactions. Triadic closure has been treated as the major mechanism for creating social links: if Alice follows Bob and Bob follows Charlie, Alice will follow Charlie. Here we present an analysis of longitudinal micro-blogging data, revealing a more nuanced view of the strategies employed by users when expanding their social circles. While the network structure affects the spread of information among users, the network is in turn shaped by this communication activity. This suggests a link creation mechanism whereby Alice is more likely to follow Charlie after seeing many messages by Charlie. We characterize users with a set of parameters associated with different link creation strategies, estimated by a Maximum-Likelihood approach. Triadic closure does have a strong effect on link formation, but shortcuts based on traffic are another key factor in interpreting network evolution. However, individual strategies for following other users are highly heterogeneous. Link creation behaviors can be summarized by classifying users in different categories with distinct structural and behavioral characteristics. Users who are popular, active, and influential tend to create traffic-based shortcuts, making the information diffusion process more efficient in the network.

Categories and Subject Descriptors

H.1 [**Information Systems**]: Models and Principles—*Systems and Information Theory*; J.4 [**Computing Applications**]: Social and Behavioral Sciences—*Sociology*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human factors, Human information processing*

Keywords

Link creation, traffic, network evolution, information diffusion, shortcut, user behavior, social media, network structure

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

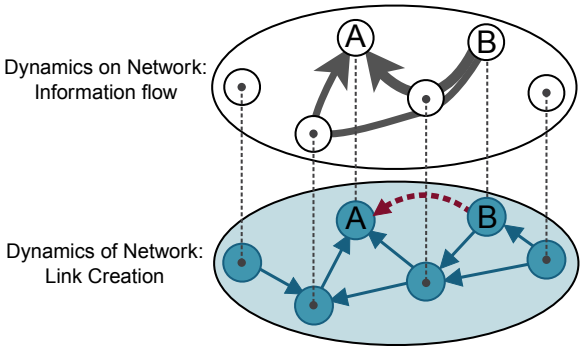

Figure 1: The dynamics *of* and *on* the network are strongly coupled. The bottom layer illustrates the social network structure, where the blue arrows represent "follow" relationships with the direction of information flow. The dashed red arrow marks a newly created link. The upper layer depicts the flow of information between people in the same group, leading to the creation of the new link.

1. INTRODUCTION

User activity in online social networks is exploding. Social and micro-blogging networks such as Facebook, Twitter, and Google Plus every day host the information sharing activity of billions of users. Using these systems, people communicate ideas, opinions, videos, and photos among their circles of friends and followers across the world. These interactions generate an unprecedented amount of data that can be used as a social observatory, providing a unique opportunity to shed light on the mechanisms of human communication with a quantitative approach [38, 15, 36, 27, 59, 60].

Research on social media revolves around two main themes: communication and its social network substrate. Most network models focus on either the structural growth of the system — the dynamics *of* the network — or information diffusion processes — the dynamics *on* the network. The present work establishes a feedback loop between these two dynamics.

Much effort has been devoted to modeling the evolution of social networks [62, 9, 46, 27]. Among proposed mechanisms of how a link is created, *triadic closure* [58, 28] is a simple but powerful principle to model the evolution of social networks based on shared friends: two individuals with mutual friends have a higher than random chance to establish a link. In directed networks, such as

Twitter or Yahoo! Meme, triadic closure implies a particular order with respect to the direction of links: once Alice follows Bob and Bob follows Charlie, Alice will follow Charlie. Triadic closure has been observed in both undirected and directed online social networks and incorporated into several network growth models [39, 34, 54]. However, most existing models do not take user activity — or how information spreads on the network — into consideration.

Social micro-blogging networks, such as Twitter, Google Plus, Sina Weibo, and Yahoo! Meme, are designed for information sharing. As illustrated in Fig. 1, the social network structure constrains communication patterns, but information propagated through the network also affect how agents behave and ultimately how the network changes and grows. In this paper we study the role of information diffusion in shaping the evolution of the network structure, and the individual strategies that bring about this effect by way of creating social links.

The major contribution of this paper is to present clear evidence that information diffusion affects network evolution at both system-wide and individual levels. In particular, we find that a considerable portion of new links are *shortcuts based on information flow* (§ 4). There is significant statistical evidence for triadic closure as a link creation mechanism, but also that users tend to link to people who have generated content they have seen (§ 4.1). Furthermore, not all users apply the same strategy to grow their social connections; users with high in-degree tend to pay more attention to traffic (§ 4.2). However, shortcuts are not equally probable; we find that users follow the most active sources of content; purely topological mechanisms cannot account for these shortcuts (§ 4.3). As a result, traffic-based shortcuts can make the social network more efficient in terms of information diffusion (§ 4.4). In § 5, we perform a Maximum Likelihood Estimation analysis to quantify the system-wide prevalence of different link creation strategies. Finally, the categorization of users suggests the existence of several distinct link formation behaviors (§ 6). Our findings identify information diffusion dynamics as a key factor in the evolution of social networks.

2. BACKGROUND

Early models concerning communication dynamics were inspired by studies of epidemics, assuming that a piece of information could pass from one individual to another through social contacts [52, 25, 18, 4, 2]. These models have been extended to include cascade phenomena [26], factors that influence the speed of spreading such as information recency [43], the heterogeneity in connectivity patterns [49], clustering [47], user-created content [6], and temporal connectivity patterns [44, 11, 12, 51]. An alternative class of models is based on the idea of a threshold; you propagate an idea when some number of friends communicate it to you [29, 45]. These models are believed to be relevant in the diffusion of rumors, norms, and behaviors [14], and have been extended to study the role of competition for finite attention [64]. The large majority of these studies consider either a *static* or *annealed* underlying social network, under the assumption that the network evolves on a longer (slower) time scale than the information spread. Recent research has addressed the modeling of intermediate cases, in which the two time scales are comparable. These approaches consider the two dynamics as either independent [53, 50] or coupled [61, 57]. The foundations of this last class of models are very similar to those explored in this paper. However, thus far, these models have focused mainly on epidemic processes in which links are deleted or rewired according to the disease status of each node [61, 57]. The social systems considered in this paper are governed by quite different underlying mechanisms.

Models devoted to reproducing the growth and evolution of network topology have traditionally focused on defining basic mechanisms driving link creation [62, 46, 9]. From the first model proposed in 1959 by Erdös and Rényi [22], many others have been introduced capturing different properties observed in real networks, such as the small-world phenomenon [63, 39, 34, 54], large clustering coefficient [63, 39, 34, 54], temporal dynamics [51, 53], information propagation [8], and heterogeneous distributions in connectivity patterns [7, 33, 37, 35, 20, 23]. In particular, this latter property was first described by the preferential attachment [7] and copy models [37].

In the social context, the rationale behind preferential attachment mechanisms is that people prefer linking to well connected individuals [7]. Although very popular, this prescription alone is not sufficient to reproduce other important features of social networks. Other models have been put forth to fill this gap, including ingredients such as homophily [31, 42, 48, 1, 24] and triadic closure [58, 28, 39, 34, 54]. Homophily describes the tendency of people to connect with others sharing similar features [42, 31]. Its impact on link creation in large-scale online networks is a recent topic of discussion [48, 1, 24]. The triadic closure mechanism is based on the intuition that two individuals with mutual friends have a higher probability to establish a link [58, 28]. This tendency has been observed in both undirected and directed online social networks and incorporated into several network growth models [39, 34, 54]. In particular Leskovec *et al.* have tested triadic closure against many other mechanisms in four different large-scale social networks [39]. By using Maximum Likelihood Estimation (MLE) [17] they have identified triadic closure as the best rule, among those considered, to explain link creation.

Link prediction algorithms, aimed at inferring new connection that may happen in the near future given a current snapshot of the network structure, could be used as ingredients for modeling network evolution. Common approaches consider link prediction as a classification task or ranking problem, using node similarity [40, 32], the hierarchical structure of the network [16], random walks [3], graphical models [41], and user profile features [56].

Although similar in spirit, our approach is different from this large body of literature. We do not consider link prediction, nor agent based simulations in which the structural behavior of each user is modeled by a set of rules. We adopt the MLE framework extending the work of Leskovec *et al.* [39]. We extend the notion of triadic closure by considering mechanisms based on traffic, or more in general, users activity. We explicitly study the coupling between the dynamics *of* and *on* the network, connecting these two previously separated themes of research, in the context of online social networks.

3. MEME DATESET

We study *Yahoo! Meme,* a social micro-blogging system similar to Twitter, which was active between 2009 and 2012. We have access to the entire history of the system, including full records of every message propagation and link creation event, from April 2009 until March 2010. A user j following a user i is represented in the follower network by a directed edge $\ell = (i, j)$, indicating j can receive messages posted by i. We adopt this notation, in which the link creator is the target, to emphasize the direction of information flow. Edges are directed to account for asymmetric relations between users; a node can follow another without being followed back. In our notation, the in-degree of a node i is the number of people followed by i, and the out-degree is the number of i's followers. Users can repost received messages, which become visible to their followers. When user j reposts content from i, we infer

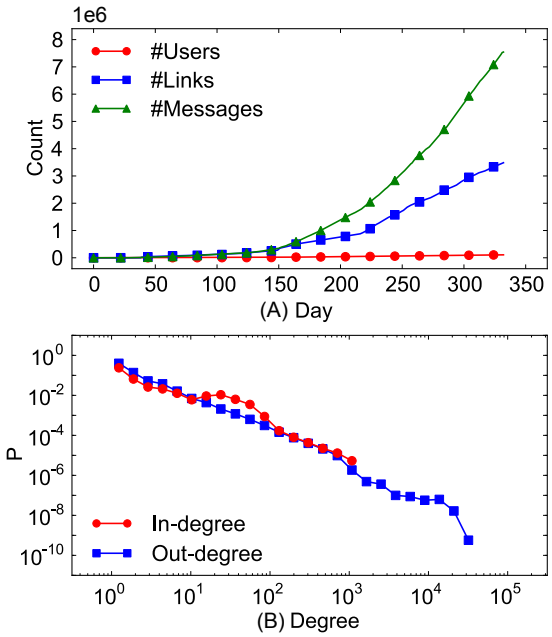

Figure 2: General statistics of the Yahoo! Meme system. (A) The growth of the system in time, the number of users (red circles), links (blue squares) and messages (green triangles). (B) Broad distributions of in-degree and out-degree in the follower network of Yahoo! Meme. Users were not allowed to follow more than 1,000 people, which is the maximum in-degree a node can attain.

a flow of information from i to j. Each link is weighted by the numbers of messages from i that are reposted or seen by j.

At the end of the observation period, the Yahoo! Meme follower network consisted of 128,199 users with at least one edge, connected by a total of 3,485,361 directed edges. Fig. 2 displays general statistics about the growth and structure of the network.

4. LINK CREATION MECHANISMS

When users post or repost messages, all their followers can see these posts and might decide to repost them, generating paths that together form cascade networks. When receiving a reposted message, a user in such a path can see both the *grandparent* (G, the user two steps ahead in the path) and the *origin* (O, original source). A user may decide to follow a grandparent or origin, receiving their future messages directly. These new links create *shortcuts* connecting users at any distance in the network. A triadic closure occurs when a user follows a *triadic node* (Δ, the user two steps away in the follower network). The definitions of different types of link creation mechanisms are illustrated in Fig. 3(A).

The Venn diagram in Fig. 3(B) shows the proportions of links of different types and the logical relationships between these sets of links. We observe that 84.8% of new edges consist of triadic closures, 21.5% form shortcuts to grandparent, and 19.5% to origins. Note that not all the grandparents are triadic nodes, because users are allowed to repost messages from people they are not following in Yahoo! Meme. This account for 0.03% of links. There is a large overlap between triadic closure links and traffic-based shortcuts. This can be explained by the phenomenon that most real-world information cascades are shallow [5] and thus triadic closure links and traffic-based shortcuts coincide.

This evidence suggests that traffic-based link creation mechanisms are an important complement to the triadic closure in modeling network evolution. Actions of posting and reposting induce the creation of shortcuts, shaping the structure of the network. Newly created links in turn determine what messages are seen by users, making the network more efficient at spreading information.

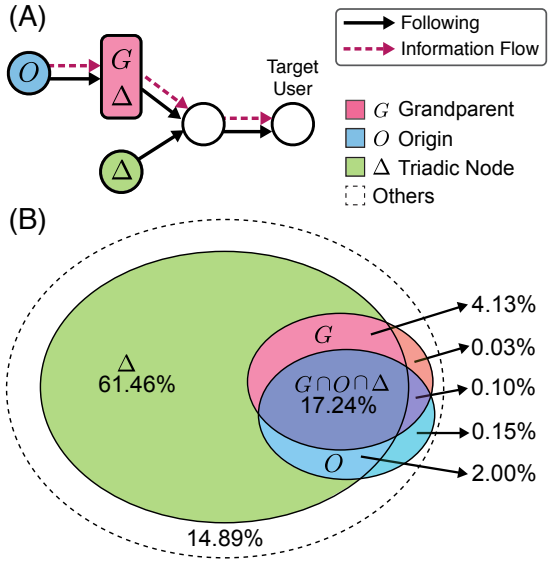

Figure 3: (A) Illustration of the link creation mechanisms. (B) Venn digram of the proportions of grandparent, origin, and triadic closure links among all existing edges.

4.1 Statistical Analyses of Shortcuts

To quantify the statistical tendency of users to create shortcuts, let us consider every single link creation in the data as an independent event. We test the null hypotheses that links to grandparents, origins, and triadic nodes are generated by choosing targets at random among the users not already followed by the creator.

We label each link ℓ by its creation order, $1 \leq \ell \leq L$, where L is the total number of links. For each link, we can compute the likelihood of following a grandparent by chance:

$$p_G(\ell) = \frac{N_G(\ell)}{N(\ell) - k(\ell) - 1}$$

where $N_G(\ell)$ is the number of distinct grandparents seen by the creator of ℓ at the moment when ℓ is about to be created; $N(\ell)$ is the number of available users in the system when ℓ is to be created; $k(\ell)$ is the in-degree of ℓ's creator at the same moment; and the denominator is the number of potential candidates to be followed. The indicator function for each link ℓ denotes whether the link connects with a grandparent or not in the real data:

$$\mathbf{1}_G(\ell) = \begin{cases} 1 & \text{if } \ell \text{ links to a grandparent} \\ 0 & \text{otherwise.} \end{cases}$$

The expected number of links to grandparents according to the null hypothesis can be then computed as:

$$E_G = \sum_{\ell=1}^{L} p_G(\ell)$$

and its variance is given by:

$$\sigma_G^2 = \sum_{\ell=1}^{L} p_G(\ell)\left(1 - p_G(\ell)\right)$$

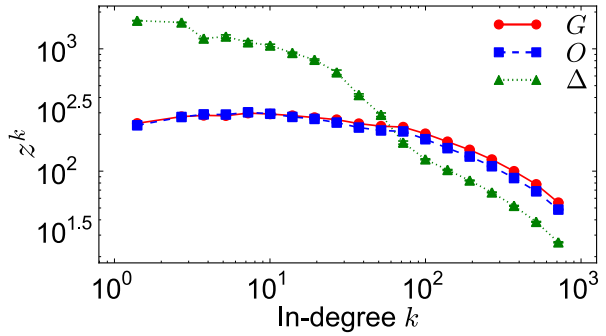

Figure 4: Individual preferences for following grandparents (red circles), origins (blue squares) and triadic nodes (green triangles) change with the in-degree of the link creator.

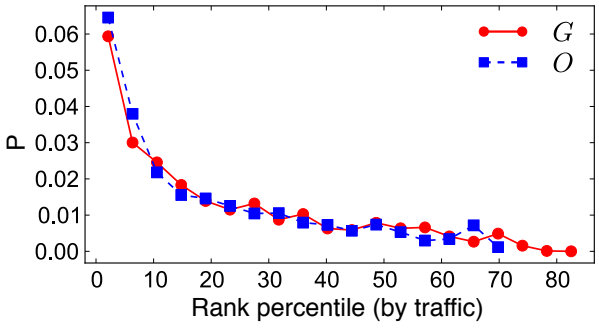

Figure 5: Probability density of followed grandparents (red circles) or origins (blue squares) having a certain rank percentile. Link targets are ranked so that the link creator has seen more messages from a user with smaller rank percentile.

while the corresponding empirical number is:

$$S_G = \sum_{\ell=1}^{L} \mathbf{1}_G(\ell).$$

According to the Lyapunov central limit theorem,[1] the variable $z_G = (S_G - E_G)/\sigma_G$ is distributed according to a standard normal $\mathcal{N}(0,1)$. For linking to origins (O) or triadic nodes (Δ), we can define z_O and z_Δ similarly. In all three cases, using a z-test, we can reject the null hypotheses with high confidence ($p < 10^{-10}$). We conclude that links established by following grandparents, origins or triadic nodes happen much more frequently than by random connection. These link creation mechanisms have important roles in the evolution of the social network.

4.2 User Preference

To study the dependence of the link formation tendencies on the different stages of an individual's lifetime, let us compute z_G^k, z_O^k and z_Δ^k for links created by users with in-degree k, that is, those who are following k users at the time when the link is created. Fig. 4 shows that the principle of triadic closure dominates user behavior when one follows a small number of users ($k < 75$). In the early stages, one does not receive much traffic, so it is natural to follow people based on local social circles, consistently with triadic closure. However, users who have been active for a long time and have followed many people ($k > 75$) have more channels through which they monitor traffic. This creates an opportunity to follow others from whom they have seen messages in the past.

4.3 Traffic Bias

Further inspection of the empirical data reveals that not all shortcuts are equally likely; users tend to follow those who have often been sources of seen messages. To investigate this, consider all new shortcuts to grandparents or origins. For each shortcut, we rank all the available grandparent or origin candidates according to how many of their messages have been seen by the creator prior to the link formation. We plot the probability of a followed grandparent or origin having a certain rank percentile in Fig. 5. The plot clearly demonstrates that repeated exposure to contents posted by a user increases the probability of following that user. This is analogous to the way in which we are more likely to adopt a piece of information or behavior to which we are exposed multiple times [6, 14, 55, 30]. This observation shows that topology alone is insufficient to explain the evolution of the network; activity patterns — the dynamics *on* the network — are a necessary ingredient in describing the formation of new links.

[1]Lyapunov's condition, $\frac{1}{\sigma_n^4}\sum_{\ell=1}^{n} E[(X(\ell) - p(\ell))^4] \overset{n\to\infty}{\longrightarrow} 0$ where $X(\ell)$ is a random Bernoulli variable with success probability $p(\ell)$ [10], is consistent with numerical tests. Details are omitted for brevity.

4.4 Link Efficiency

In information diffusion networks like Twitter and Yahoo! Meme, social links may have a key efficiency function of shortening the distance between information creators and consumers. An efficient link should be able to convey more information to the follower than others. Hence we define the *efficiency* of link ℓ as the average number of posts seen or reposted through ℓ during one time unit after its creation:

$$\eta_{\text{seen}} = \frac{w_{\text{seen}}(\ell)}{T - t(\ell)}, \quad \eta_{\text{repost}} = \frac{w_{\text{repost}}(\ell)}{T - t(\ell)}$$

where $w(\ell)$ is the number of messages seen or reposted through ℓ; $t(\ell)$ is the time when ℓ was created; and T is the time of the last action recorded in our dataset. Both seen and reposted messages are considered, as they represent different types of traffic; the former are what is visible to a user, and the latter are what a user is willing to share. We compute the link efficiency of every grandparent, origin, and triadic closure link. As shown in Fig. 6, both grandparent and origin links exhibit higher efficiency than triadic closure links, irrespective of the type of traffic. By shortening the paths of information flows, more posts from the content generators reach the consumers.

5. RULES OF NETWORK EVOLUTION

To infer the different link creation strategies from the observed data, we characterize users with a set of probabilities associated with different actions, and approximate these parameters by Maximum-Likelihood Estimation (MLE) [17]. For each link ℓ, we know the actual creator and the target; we can thus compute the likelihood $f(\ell|\Gamma,\Theta)$ of the target being followed by the creator according to a particular strategy Γ, given the network configuration Θ at the time when ℓ is created. The likelihoods associated with different strategies can be mixed according to the parameters to obtain a model of link creation behavior. Finally, assuming that link creation events are independent, we can derive the likelihood of obtaining the empirical network from the model by the product of likelihoods associated with every link. The higher the value of the likelihood function, the more *accurate* the model.

5.1 Single Strategies

Let us consider five link creation mechanisms and their combinations:

Random (Rand): follow a randomly selected user who is not yet followed.

Triadic closure (Δ): follow a randomly selected triadic node.

Grandparent (G): follow a randomly selected grandparent.

Origin (O): follow a randomly selected origin.

Traffic shortcut ($G \cup O$): follow a randomly selected grandparent or origin.

Other mechanisms for link creation could be similarly incorporated, such as social balance [21] and preferential attachment [7]. However, preferential attachment is built on the assumption that everyone knows the global connectivity of everyone else, which is not realistic. The strategies considered here essentially reproduce and extend the copy model [37], approximating preferential attachment with only local knowledge.

To model link creation with a single strategy, we can use a parameter p for the probability of using that strategy, while a random user is followed with probability $1-p$. The calculation of maximum likelihood, taking the single strategy of grandparents as an example, is as follows:

$$\begin{aligned}\mathcal{L}_G(p) &= \prod_{\ell=1}^{L} \left(pf(\ell|G,\Theta) + (1-p)f(\ell|\text{Rand},\Theta)\right) \\ &= \prod_{\ell=1}^{L} \left(p\frac{\mathbf{1}_G(\ell)}{N_G(\ell)} + (1-p)\frac{1}{N(\ell)-k(\ell)-1}\right) \\ &= \prod_{\mathbf{1}_G(\ell)=1} \left(\frac{p}{N_G(\ell)} + \frac{1-p}{N(\ell)-k(\ell)-1}\right) \\ &\quad \prod_{\mathbf{1}_G(\ell)=0} \frac{1-p}{N(\ell)-k(\ell)-1}.\end{aligned}$$

Note that since a follow action can be ascribed to multiple strategies, it can contribute to multiple terms in the log-likelihood expression. For instance, a link could be counted in both $f(\ell|G,\Theta)$ and $f(\ell|\text{Rand},\Theta)$. For numerically stable computation, we maximize the log-likelihood:

$$\log \mathcal{L}_G(p) = \sum_{\mathbf{1}_G(\ell)=1} \ln\left(\frac{p}{N_G(\ell)} + \frac{1-p}{N(\ell)-k(\ell)-1}\right) + \sum_{\mathbf{1}_G(\ell)=0} \ln \frac{1-p}{N(\ell)-k(\ell)-1}.$$

Similar expressions of log-likelihood can be obtained for other strategies (Δ, O, and $G \cup O$).

It is not trivial to obtain the best p analytically, so we explore the values of $p \in (0,1)$ numerically (Fig. 7). Triadic closure dominates as a single strategy, with $p_\Delta = 0.82$, consistently with the large number of triadic closure links observed in the data. Traffic-based strategies alone account for about 20% of the links.

5.2 Combined Strategies

For a more realistic model of the empirical data, let us consider combined strategies with both triadic closure and traffic-based shortcuts. For each link ℓ, the follower with probability p_1 creates a shortcut by linking to a grandparent (G), an origin (O), or either of them ($G \cup O$); with probability p_2 follows a triadic node (Δ); and with probability $1-p_1-p_2$ connects to a random node.

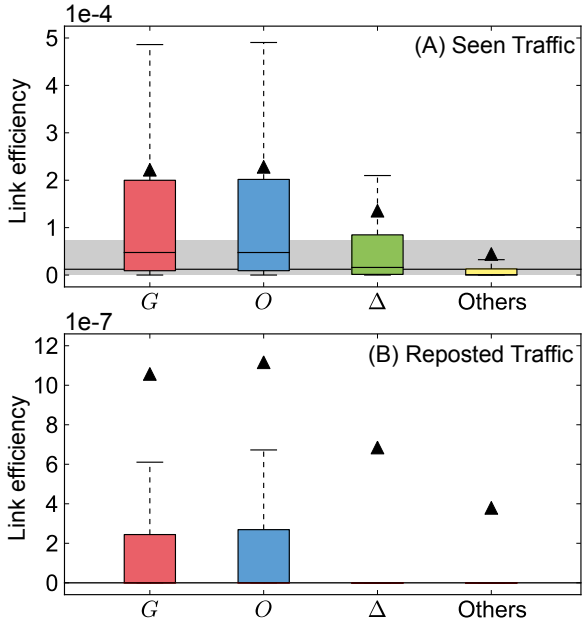

Figure 6: Efficiency of links created according to different mechanisms, or average number of messages (A) seen or (B) reposted per time unit. Each box shows data within lower and upper quartile. Whiskers represent the 99th percentile. The triangle and line in a box represent the median and mean, respectively. Note that the mean can fall outside the shown quantiles for skewed distributions. The grey area and the black line across the entire figure mark the interquartile range and the median of the measure across all links, respectively.

Taking the combined strategy with grandparent as an example, we compute the log-likelihood as:

$$\begin{aligned}&\log \mathcal{L}_{G+\Delta}(p_1,p_2) \\ &= \log \prod_{\ell=1}^{L} [p_1 f(\ell|G,\Theta) + p_2 f(\ell|\Delta,\Theta) \\ &\quad + (1-p_1-p_2) f(\ell|\text{Rand},\Theta)] \\ &= \sum_{\substack{\mathbf{1}_G(\ell)=1 \\ \mathbf{1}_\Delta(\ell)=1}} \log\left(\frac{p_1}{N_G(\ell)} + \frac{p_2}{N_\Delta(\ell)} + \frac{1-p_1-p_2}{N(\ell)-k(\ell)-1}\right) \\ &\quad + \sum_{\substack{\mathbf{1}_G(\ell)=1 \\ \mathbf{1}_\Delta(\ell)=0}} \log\left(\frac{p_1}{N_G(\ell)} + \frac{1-p_1-p_2}{N(\ell)-k(\ell)-1}\right) \\ &\quad + \sum_{\substack{\mathbf{1}_G(\ell)=0 \\ \mathbf{1}_\Delta(\ell)=1}} \log\left(\frac{p_2}{N_\Delta(\ell)} + \frac{1-p_1-p_2}{N(\ell)-k(\ell)-1}\right) \\ &\quad + \sum_{\substack{\mathbf{1}_G(\ell)=0 \\ \mathbf{1}_\Delta(\ell)=0}} \log \frac{1-p_1-p_2}{N(\ell)-k(\ell)-1}.\end{aligned}$$

Once again, many follow actions can create both triadic closure links and traffic shortcuts, so they can contribute to multiple terms in the log-likelihood expression.

It is hard to obtain the optimal solution analytically. We numerically explore the values of p_1 and p_2 in the unit square to maximize the log-likelihood. The best combined strategy is the one consid-

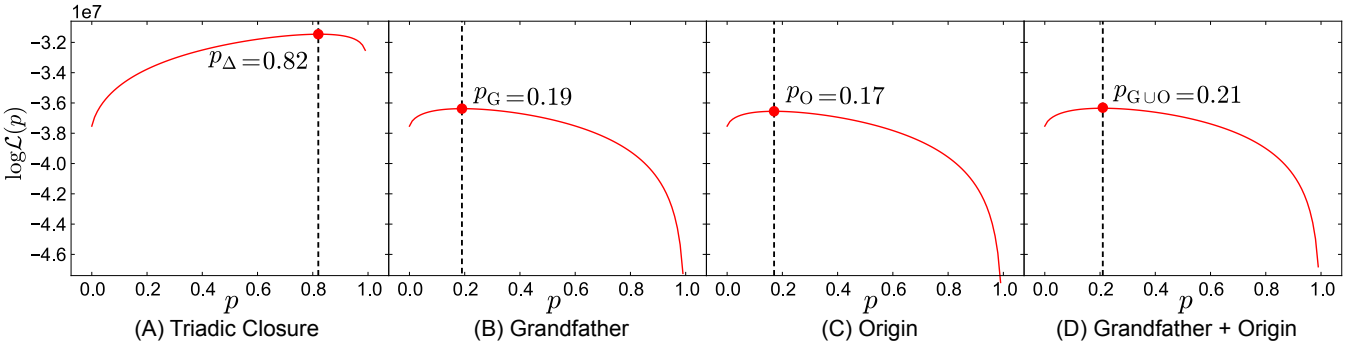

Figure 7: Plot of the log-likelihood $\log \mathcal{L}(p)$ as a function of link creation strategy probabilities for models with a single strategy. The red circles mark the maximized $\log \mathcal{L}(p)$.

ering both grandparents and origins as well as triadic closure (see Fig. 8). The parameter settings and the maximum likelihood values for all tested models are listed in Table 1. We can compare the quality of these models by comparing their maximized $\log \mathcal{L}$'s. The combined models with both traffic shortcuts and triadic closure yield the best accuracy. In these models, triadic closure accounts for 71% of the links, grandparents and origins for 12%, and the rest are created at random.

Thus far we have assumed that each user has the same behavior; in the next section we model each user separately.

6. USER BEHAVIOR

The MLE models for describing the system behavior can be similarly employed to characterize the strategy of an individual user. Let us focus on the model $G \cup O + \Delta$ that best reproduces the empirical data at the global level. We run MLE to explain the links created by each user independently. We consider users with at least 20 in-links, such that MLE is meaningful. For easier interpretation, let us call $p_{\text{traffic}} = p_1$, $p_{\text{structure}} = p_2$ and $p_{\text{random}} = 1 - p_1 - p_2$. Each user has her own set of parameters.

6.1 User Strategy Classification

Using the Expectation-Maximization (EM) algorithm [13, 19], users are clustered into several classes based on p_{traffic}, $p_{\text{structure}}$ and p_{random}. EM iteratively preforms an expectation step to compute the probability that each instance belongs to each class, and a maximization step in which latent variables of classes are altered to maximize the expected likelihood of the observed data. EM decides how many clusters to create by cross validation. This procedure yields five classes:

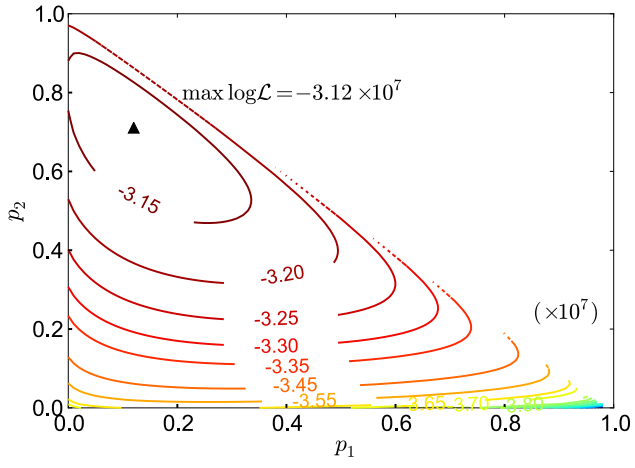

Figure 8: The contour plot of log-likelihood $\log \mathcal{L}(p_1, p_2)$ for the combined strategy of creating traffic shortcuts ($G \cup O$) with probability p_1 and triadic closure links (Δ) with probability p_2. The black triangle marks the optimum.

Table 1: The best parameters in different models and corresponding values of maximized log-likelihood function.

Strategy	Model	Parameters	max $\log \mathcal{L}$
Baseline	Rand	–	-3.75×10^7
Single	Δ	$p = 0.82$	-3.15×10^7
	G	$p = 0.19$	-3.64×10^7
	O	$p = 0.17$	-3.65×10^7
	$G \cup$ O	$p = 0.21$	-3.63×10^7
Combined	$G + \Delta$	$p_1 = 0.12$ $p_2 = 0.71$	-3.12×10^7
	O $+ \Delta$	$p_1 = 0.10$ $p_2 = 0.73$	-3.13×10^7
	$G \cup$ O $+ \Delta$	$p_1 = 0.12$ $p_2 = 0.71$	-3.12×10^7

Information-Oriented (Info): People prefer to follow someone from whom or through whom they have received messages.

Friend of a Friend (Friend): People follow users two steps away to form triadic closure, almost exclusively.

Casual Friendship (CFrd): People tend to follow a set of users their friends are following; they also link to random users occasionally.

Mixture (Mix): Miscellaneous behavior of creating traffic shortcuts, connecting others by triadic closure, and following random people.

Random Browsing (Rand): People have a much higher preference for following a random user who is not close in either the follower or the message flow network. "Random" does not necessarily imply the absence of any rule; there can be other strategies not explored in our model, i.e., following a celebrity on purpose (similar to preferential attachment).

Table 2 displays the parameter averages for users in each class, representing the overall behavior pattern in that class. Users in the mixture category behave similarly to the average across all users. Fig. 9 illustrates how users in different classes are mapped into the parameter space with the probability of each link creation strategy as one dimension.

Table 2: Classes of user link creation strategy

Class	#Users	$\langle p_{\text{traffic}} \rangle$	$\langle p_{\text{structure}} \rangle$	$\langle p_{\text{random}} \rangle$
All Users	45,708	0.07	0.77	0.17
Info	4,750	0.52	0.36	0.13
Friend	12,797	0.00	0.96	0.04
CFrd	23,469	0.01	0.80	0.19
Mix	2,524	0.07	0.63	0.30
Rand	2,168	0.09	0.32	0.59

6.2 Characterization of User Classes

To further differentiate users with different link creation strategies, let us look at several structural and behavioral characteristics of each class. Figs. 10(A-C) show how users in different classes create social links by comparing p_{traffic}, $p_{\text{structure}}$ and p_{random}.

As shown in Fig. 10(D), information-oriented users have been active longer than users in friendship classes. Similarly, information-oriented users tend to follow more people (Fig. 10(E)). Information-oriented users have even more followers compared to friendship-driven users (Fig. 10(F)). This suggests that they tend to be more influential, as confirmed by considering the number of times that their messages are reposted (Fig. 10(G)). Friendship-driven users follow a few people while essentially nobody is following them. Such a passive role can be explained by their short lifetime. All of these results are consistent with Fig. 4.

Finally, Figs. 10(H-I) suggest that, while information-oriented users tend to produce more messages, their role is more that of spreaders than producers of information compared to other classes.

In this analysis, the parameters are fit according to the entire lifetime of each user. Focusing on the first 20 or 50 links does not yield qualitatively different results.

7. CONCLUSION

The study of the feedback loop between the dynamics *of* and *on* the network — how the network grows and how the information spreads — offers a promising framework for understanding social influence, user behavior, and network efficiency in the context of micro-blogging systems.

The results presented in this paper show that while triadic closure is the dominant mechanism for social network evolution, it is mainly relevant in the early stages of a user's lifetime. As time progresses, the traffic generated by the dynamics of information flow on the network becomes an indispensable component for user linking behavior. As users become more active and influential, their links create shortcuts that make the spread of information more efficient in the network. Users whose following behavior is driven by the information they see are a minority of the population, but play a key role in the information diffusion process. They produce more information, but, even more importantly, they act as spreaders of the information they collect widely across the network.

While existing link prediction algorithms [40, 16, 3, 41, 56] are not designed to explain the network evolution in a dynamic setting, the MLE framework could in principle be used to assess which link prediction methods are more consistent with the longitudinal structural changes observed in the network, by treating the prediction at each step as a link creation strategy. These approaches will be explored in future work.

We believe our findings apply generally to techno-social networks, and in particular information diffusion networks and (micro)blogs. Analyses of other micro-blogging systems, such as Twitter and Weibo, would be needed to confirm this, but will be challenging due to the difficulty of obtaining full longitudinal data about user actions on the social network.

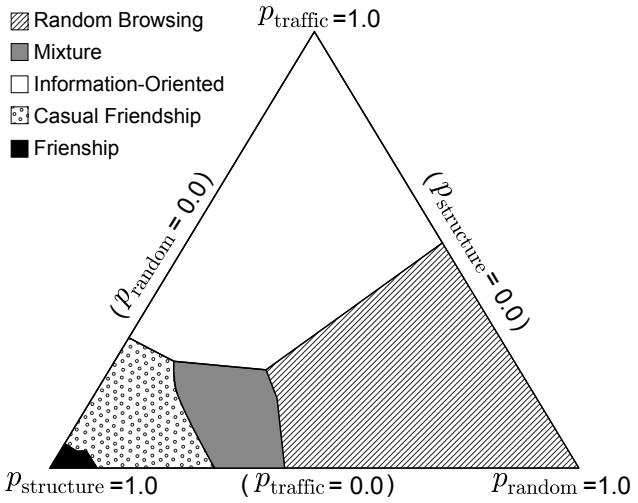

Figure 9: Ternary plot of users according to p_{traffic}, $p_{\text{structure}}$ and p_{traffic}.

Acknowledgments. This project is a collaboration between the Center for Complex Networks and Systems Research (`cents.indiana.edu`) at Indiana Univ. and Yahoo! Research Barcelona. During the project, CC was at Yahoo!; NP and BG were at CNetS; RS was visiting CNetS; and JR was visiting Yahoo! as a PhD student at CNetS. We are grateful to Alessandro Vespignani for helpful discussion, to Ricardo Baeza-Yates for support and access to the data, and to anonymous referees for valuable comments. This project was funded in part by the James S. McDonnell Foundation, NSF Grant No. 1101743, DARPA Grant No. W911NF-12-1-0037, and the Indiana Univ. School of Informatics and Computing.

8. REFERENCES

[1] L. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. *ACM Trans. Web*, 6(2):9, 2012.

[2] R. M. Anderson and R. M. May. *Infectious Diseases in Humans*. Oxford Univ. Press, 1992.

[3] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*, pages 635–644. ACM, 2011.

[4] N. Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London, 2nd edition, 1975.

[5] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*, pages 65–74. ACM, 2011.

[6] E. Bakshy, B. Karrer, and L. Adamic. Social influence and the diffusion of user created content. In *Proc. ACM Conf. on Electronic Commerce*, 2009.

[7] A.-L. Barabási and R. Albert. Emergence of scaling in random graphs. *Science*, 286:509–511, 1999.

[8] N. Barbieri, F. Bonchi, and G. Manco. Cascade-based community detection. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*, pages 33–42, 2013.

[9] A. Barrat, M. Barthèlemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

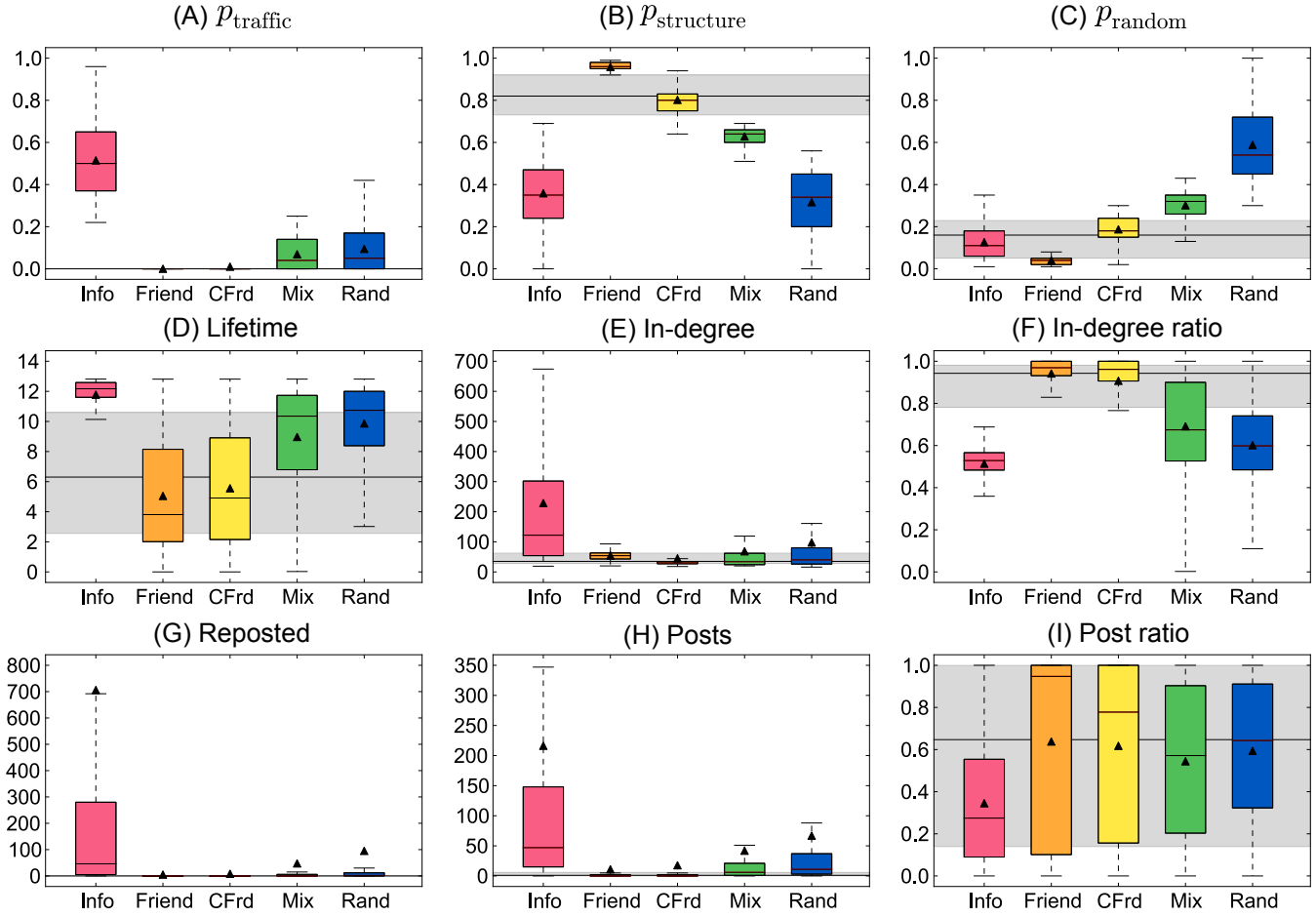

Figure 10: Various features of users in different classes. The lifetime of a user is measured by how many others join the system after him. The in-degree is the number of people a user is following, k, and the in-degree ratio is $k/(k + k_{out})$ where k_{out} is the number of followers. "Reposted" refers to the number of times that a user's messages are reposted by others. "Posts" denotes the number of messages generated by a user excluding reposts. The post ratio is the fraction of all posts by a user (including reposts) that are originated by that user. Each box shows data within lower and upper quartile. Whiskers represent the 99th percentile. The triangle and line in a box represent the median and mean, respectively. The grey area and the black line across the entire figure mark the interquartile range and the median of the measure across all links, respectively.

[10] P. Billingsley. *Probability and measure*, page 362. John Wiley & Sons, 1995.

[11] C. Butts. Relational event framework for social action. *Sociological Methodology*, 38:155–200, 2008.

[12] C. Butts. Revisting the foundations of network analysis. *Science*, 325:414–416, 2009.

[13] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332, 1992.

[14] D. Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.

[15] A. Cho. Ourselves and our interactions: the ultimate physics problem? *Science*, 325:406, 2009.

[16] A. Clauset, C. Moore, and M. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

[17] G. Cowan. *Statistical Data Analysis*. Oxford Science Publications, 1998.

[18] D. J. Daley and D. G. Kendall. Epidemics and rumours. *Nature*, 204(4963):1118, 1964.

[19] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Soc. B (Methodological)*, pages 1–38, 1977.

[20] S. Dorogovtsev, J. Mendes, and A. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, 85:4633–4636, 2000.

[21] D. Easley and J. Kleinberg. *Networks, crowds, and markets*. Cambridge Univ Press, 2010.

[22] P. Erdös and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 5:17–61, 1960.

[23] S. Fortunato, A. Flammini, and F. Menczer. Scale-free network growth by ranking. *Phys. Rev. Lett.*, 96(21):218701, 2006.

[24] L. Gallos, D. Rybski, F. Liljeros, S. Havlin, and H. Makse. How people interact in evolving online affiliation networks. *Phys. Rev. X*, 2(3):031014, 2012.

[25] W. Goffman and V. A. Newill. Generalization of epidemic theory: an application to the transmission of ideas. *Nature*, 204(4955):225–228, 1964.

[26] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 3(12):211–223, 2001.

[27] B. Goncalves, N. Perra, and A. Vespignani. Modeling users' activity on Twitter networks: Validation of Dunbar's number. *PLoS ONE*, 6(8):e22656, 2011.

[28] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.

[29] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1433, 1978.

[30] N. Hodas and K. Lerman. How limited visibility and divided attention constrain social contagion. In *Proc. ASE/IEEE Intl. Conf. on Social Computing (SocialComm)*, 2012.

[31] P. Holme and M. E. J. Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Phys. Rev. E*, 74(5), 2006.

[32] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):38–43, 1953.

[33] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: measurements, models and methods. *Lecture Notes in Computer Science (LNCS)*, 1627:1–18, 1999.

[34] D. Krackhardt and M. Handcock. Heider vs. Simmel: Emergent features in dynamic structure. *Statistical Network Analysis: Models, Issues, and New Directions*, 4503:14–27, 2007.

[35] P. Krapivsky and S. Redner. Organization of growing random networks. *Phys. Rev. E*, 63:066123, 2001.

[36] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proc. SIGKDD Intl. Conf. on Knowledg discovery and data mining*, 2006.

[37] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. Annual Symposium on Foundations of Computer Scienc*, pages 57–65. IEEE, 2000.

[38] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.

[39] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proc. SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pages 462–470, 2008.

[40] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *J. American Society for Info. Sci. and Tech.*, 58(7):1019–1031, 2007.

[41] T. Lou, J. Tang, J. Hopcroft, Z. Fang, and X. Ding. Learning to predict reciprocity and triadic closure in social networks. *ACM Trans. on Embedded Computing Systems*, 9(4), 2010.

[42] M. McPherson, L. Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[43] Y. Moreno, M. Nekovee, and A. Vespignani. Efficiency and reliability of epidemic data dissemination in complex networks. *Phys. Rev. E*, 69:055101(R), 2004.

[44] M. Morris and M. Kretzschmar. Concurrent partnerships and transmission dynamics in networks. *Social Networks*, 17:299, 1995.

[45] S. Morris. Contagion. *Review of Economic Studies*, 67(1):57–78, 2000.

[46] Newman, M.E.J. *Networks, an Introduction*. Oxford University Press, 2010.

[47] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A.-L. Barabasi. Structure and tie strengths in mobile communication networks. *Proc. National Academy of Sciences*, 104:7332, 2007.

[48] F. Papadopoulos, M. Kitsak, M. Ángeles Serrano, M. Boguña, and D. Krioukov. Popularity versus similarity in growing networks. *Nature*, 489:537–540, 2012.

[49] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, 2001.

[50] N. Perra, A. Baronchelli, D. Mocanu, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani. Random walks and search in time varying networks. *Phys. Rev. Lett.*, 109:238701, 2012.

[51] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani. Time scales and dynamical processes in activity driven networks. *Scientific Reports*, 2:469, 2012.

[52] A. Rapoport. Spread of information through a population with socio-structural bias: I. assumption of transitivity. *Bull. Math. Biol.*, 15(523–533), 1953.

[53] L. E. C. Rocha, F. Liljeros, and P. Holme. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput Biol*, 7(3):e1001109, 03 2011.

[54] D. M. Romero and J. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. In *Proc. Intl. Conf. on Weblogs and Social Media (ICWSM)*. AAAI, 2010.

[55] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proc. Intl. Conf. on World Wide Web*, 2011.

[56] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in folksonomies: social link prediction from shared metadata. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*, pages 271–280, 2010.

[57] L. B. Shaw and I. B. Schwartz. Enhanced vaccine control of epidemics in adaptive networks. *Phys. Rev. E*, 81:046120, 2010.

[58] G. Simmel and K. H. Wolff. *The Sociology of Georg Simmel*. The Free Press, 1950.

[59] A. Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, 2009.

[60] A. Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nature Physics*, 8(1):32–39, 2011.

[61] E. Volz and L. A. Meyers. Epidemic thresholds in dynamic contact networks. *J. R. Soc. Interface*, 6:233241, 2009.

[62] S. Wasserman and K. Faust. Social network analysis. *Cambridge University Press.*, 1994.

[63] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393:440–442, 1998.

[64] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2, 2012.