Inferring Unobservable Inter-community Links in Large Social Networks

Heath Hohwald, Manuel Cebrián, Arturo Canales, Rubén Lara, and Nuria Oliver
Telefonica Research
Madrid, Spain
Email: {heath, manuelc, acg, rubenlh, nuriao}@tid.es

Abstract—**Social networks can be used to model social interactions between individuals. In many circumstances, not all interactions between individuals are observed. In such cases, a social network is constructed with the data that has been observed, as this is the best one can do. Recent research has attempted to predict future links in a social network, though this has proven a very challenging task. Rather than predicting future links, we propose an inference method for recovering the links in a social network that already exist but that have not been observed. In addition, our approach automatically identifies groups of individuals that form tight-knit communities and models the intra and inter-community interactions. At this higher level of abstraction and for a social network built from mobile phone calls, our method is able to accurately identify a subset of 10% of all community pairs where about 50% of the pairs have had unobserved communication between them, an improvement of about four times over a subset of the same size with randomly chosen pairs. To the best of our knowledge, this is the first method that infers links that exist but are unobservable in a phone call-based social network. In addition, we perform the inference at the community level, where the discovery of unobserved inter-community communication can provide further insight into the organizational structure of the social network and can identify social groups that may share common interests.**

I. Introduction

Social networks have been studied for over a century to connote complex sets of relationships among members of social systems. The availability of vast amounts of data describing explicit communication and social interactions between individuals (phone call logs, chat logs, e-mail logs, on-line social networks data, or location and presence data), not only for small groups of individuals but also for entire regions or even countries, has opened new possibilities for the study of social structures and dynamics in human societies [1].

Gaining insights into human social networks has a myriad of applications, ranging from the study of the spread of epidemics or the detection of individuals at risk of social exclusion, to the social recommendation of products and the design of viral marketing campaigns [2]. Furthermore, the detection of strongly connected social groups, or communities, has particular interest as their members exert mutual influence.

However, a fundamental challenge in this field consists of unveiling, at a large scale, the *real* social structure of a group of individuals from the observations available. Studies such as the reality mining experiment [3] have traced a wide range of activities of individuals and their social interactions over a period of time. However, such approaches are limited to small to medium-sized groups (in the hundreds), as they require specific devices and sensors to be available and worn by all participants. Therefore, most if not all the traces of social relations available for large groups of individuals (in the thousands to millions), such as e-mail logs or phone calls, constitute a partial view of all social interactions individuals are involved in and, thus, the social graph reconstructed from these data can only be regarded as an approximation of the real social network.

In this work, we address the problem of inferring non-observed interactions between social communities. In particular, given a social graph built from the phone communication between individuals, and an agglomerative community detection process, our method infers whether individuals in two communities are socially connected or not, even though their communication link has not been observed from available data. This inference enables the estimation of the communication strength between communities beyond observable communication, which can be used, for example, to determine if two communities should be merged into one single community during the community identification process.

The paper is organized as follows: Section II presents related work on detecting social communities and inferring missing links in social networks; Section III provides a detailed definition of the missing link inference problem addressed in our research; Section IV describes how the social network is built and the social communities are detected, our experimental setting, and the features extracted for inferring missing links between communities; our experimental results are presented in Section V; finally, Section VI discusses the results and outlines future work.

II. Related work

Community detection approaches can be grouped into two categories: a) approaches that allow *overlapping* communities, *i.e.,* individual nodes can belong to more than one community (*e.g.* [4], [5]); and b) approaches that require each node to belong to at most one community (*e.g.* [6], [7]). An orthogonal categorization classifies community detection algorithms into: a) *agglomerative*, which start with each node in the graph being a community and iteratively merge communities until a stopping condition is met (*e.g.* [8]), and b) *divisive*, which start considering that all nodes belong to the same, single community and iteratively partition this initial community into smaller groups until no further partitions can be made or a stopping condition is met (*e.g.* [7]).

Link inference and *link prediction* are two recent statistical machine learning problems that appear as a result of the increasing interest in the broader problem of *link mining* in social networks, for which [9] is an excellent survey. Recent studies have motivated the practical importance of link mining in the telecommunication industry [17], in ecology and the World Wide Web [18].

III. Inferring Unobservable Inter-community Links

In this section, we highlight the key research challenges in recovering unobserved interactions between communities in a social network.

A. Interaction Between Social Communities

One of the challenges in recovering unobserved links is that most real-world social graphs tend to be sparse. If a graph consists of N individuals, then the number of potential pairs of individuals that may interact with one another is $\binom{N}{2}$, which grows quadratically with the number of individuals. However, in social networks based on phone call data, the number of interactions actually realized is often seen to grow only linearly with the number of nodes. Therefore, any data driven machine learning method aimed at accurately recovering missing links will suffer from the typical problems present in an imbalanced class learning problem [19], namely having many more instances of pairs that have not communicated than instances that have communicated.

Partially motivated by graph sparsity, we posit that unobservable links should be easier to recover at the community level than at the individual level. Communities are characterized by strong intra-community communication links, such that information interchanged between two communities will eventually reach a large portion of the members of each of the communities, hence, the importance of accurate inter-community link estimation. Finally, there are applications and services that work at a community level, such that any unobserved interactions that could be recovered at that level would greatly improve the quality of service, *e.g.,* in the area of market segmentation [20].

Note that, unlike the case of modeling communication between individuals, many community finding algorithms allow for individuals to be members of multiple communities (see Section II). As will be seen, multiple community membership can be used as an input variable for learning algorithms, with the intuition that two communities that share one or more members are more likely to communicate.

When asking whether communication between communities exists, the same ambiguity arises as for individuals: what constitutes sufficient communication for linking two communities depends on the criteria used. In this work, we consider that two communities are linked when there is at least one successful phone call from a member of one community to a member of another community. This low threshold generates an upper bound on the number of unobserved inter-community links.

B. Recovering Unobserved Interactions Between Communities

In this paper, we focus on recovering missing edges that exist in a social network without estimating their associated weights. The recovery of edge weights would be a natural extension of the models we present. Since we make no attempt to recover edge weights, we can frame the recovery of missing links as a binary classification problem, where features extracted from the observed data are used to determine whether there has been (unobserved) communication between a pair of communities or not. Following the work of Liben-Nowell and Kleinberg [12], we focus on finding the subset of community pairs with the highest probability of having inter-community communication.

C. Simulating Unobserved Data from Observed Data

One of the difficulties in applying a supervised machine learning algorithm to tackle the problem of recovering missing links is that we often do not have the required labels (ground truth). In order to produce labels, we use a quasi-bootstrap process. First, we make the simplifying –yet reasonable– assumption that unobserved nodes behave similarly to observed nodes; the quality of results will depend on the extent to which this assumption is valid. The nodes in the original social network are divided into two sets: Γ, the set of nodes for which all edges are observed; and Λ, the set of nodes for which we can only observe edges that connect with a node in Γ. We assume that all nodes $\gamma_i \in \Gamma$ behave similarly to all nodes $\lambda_j \in \Lambda$.

The labels are generated by first randomly partitioning Γ into Γ_o and Γ_u such that all the nodes in Γ_o are fully observable nodes (*i.e.,* their edges are observed), and the nodes in Γ_u are those with edges connecting with a node in Γ_o. Therefore, the nodes in Γ_u simulate the nodes in Λ. The original social network is re-processed and all edges between Γ_u nodes are labeled as hidden. At the end of this process, a *partially* observed data set is generated that will be used as training data. Since edges connecting two nodes in Γ_u are originally observed –but labeled as hidden, we can quantitatively evaluate the ability to recover them. If nodes in Γ and Λ do in fact behave similarly, then the model learned by hiding information in Γ can be used to recover missing information about edges connecting nodes in Λ.

IV. Application area

A. Recovering Unobserved Inter-community Mobile Phone Communication

Our experimental scenario consists of inferring unseen communication between communities of mobile phone users, where the input is a set of CDR (Call Detail Record) data. This data consists of records of phone calls, one record per call. For the experiments described in this paper, we used CDRs from a single mobile phone carrier (from now on referred to as carrier A) from a metropolitan region for a six month period. The information used from each CDR included the encrypted originating phone number, the encrypted destination phone number, the duration of the call, the time and date of the call, and the carrier for both parties. In our experiments we restrict the data by selecting only CDRs where both the caller and receiver of the call were subscribers of carrier A.

From the original data set, we randomly selected 50% of the subscribers and labeled them as being subscribers of some other carrier (carrier B). Next, the original data set was reprocessed by removing all the calls where both callers were labeled as subscribers of carrier B. These calls will be considered *unobservable*, and our goal is to measure how accurately we can recover such calls while working at the community level. We shall refer to this new data set as the *partially observed* data set, and it serves as the input to the community building algorithm, described next.

B. Community Detection

Let $G = (V, E)$ be the undirected social graph built from the CDR data, where the set of vertices in the graph, V, is the set of unique encrypted phone numbers that appear in the CDRs; and the set of edges, E, includes links between pairs of customers that have had at least one communication (voice call) during the observation period. The goal of the community detection process is to find groups of individuals that form a cohesive social group. In such process we must allow for overlapping communities as people usually belong to more than one social group, *e.g.* family, friends, colleagues, etc. Hence, nodes in the graph G are not enforced to exclusively belong to one community as the goal is to identify groups of individuals for which there is enough evidence of a permanent and strong social interaction.

We follow an agglomerative, hierarchical process in order to identify overlapping communities in the graph. The final result of the process is a set $\Lambda = \{C_1, \ldots, C_n\}$ of communities, such that $C_i \subseteq V$, and there is a high number of social links (edges) among the members of the group[1]. Furthermore, pairs of nodes connected by an edge (dyads) that are not contained in any community $C_i \in \Lambda$ are not considered a community, as in our experiment we want to focus on communities of at least 3 members.

Given the data set previously presented and an original graph with $267,230$ nodes and $1,561,170$ links, the community detection process yields $11,162$ (overlapping) communities of sizes ranging from 3 to 11. After simulating that 50% of all numbers are subscribers of another carrier (B), we have a partially observed graph with $248,456$ nodes, $1,166,960$ links and $6,963$ (overlapping) communities. It can be thus seen that the unobservability of links in the graph has a direct impact on what communities result from the community identification process.

C. Sampling Methodology for Constructing the Training Data

In order to train our models, we build a training set where each observation is a feature vector (see next section) about a pair of communities, $(C_i, C_j), i \neq j$. Its associated target value is a binary variable $y \in \{0, 1\}$, with $y = 0$ when C_i and C_j do not communicate, while $y = 1$ when C_i and C_j do communicate.

Given N communities, there is a total of $\binom{N}{2}$ pairs of communities. Note that the vast majority of community pairs (more than 99.9%) have no communication between them, *i.e.,* the pair-wise communication matrix is extremely sparse. Therefore, we only use in our training set community pairs that have at least one observed call between the two communities. This restriction greatly reduces the number of observations in the training set and also mitigates the class imbalance problem due to the sparsity of the data [19].

D. Description of the Feature Set

We focus on features that utilize community information and that can be directly observed from the CDR data. Note that complex topological features that have been previously proposed to predict individual behaviors [12], are not easily extended to the community prediction problem. Therefore, we propose a new set of community-centric features.

In particular, we use a 20-dimensional community feature vector that includes *community size, community overlap, cross-community call frequency, cross-community call duration, intra-community call frequency,* and *intra-community call duration* features. We describe each of the features in further detail.

1) Community Size: Given a community pair (C_i, C_j), let $\phi(C_i)$ be the set of all individuals that are members of community C_i. In each community C_i, let $\phi_A(C_i)$ and $\phi_B(C_i)$ be the set of individuals of that community that are subscribers to carriers A and B, respectively. This distinction yields four community size-related features:

$$f_1 = \left|\phi_A(C_i)\right|, f_2 = \left|\phi_B(C_i)\right|$$
$$f_3 = \left|\phi_A(C_j)\right|, f_4 = \left|\phi_B(C_j)\right|$$

where $\left|\phi_j(C_i)\right|$ denotes the number of individuals that are members of community C_i and are subscribers of carrier j. Community size features should be helpful in the prediction of communication between communities. Intuitively, the larger the communities, the higher the probability of inter-community communication. Since the community discovery process includes members of both carriers, all information needed to produce these features is observable.

2) Community Overlap: Community size features capture information about the size of each community. However, they do not include any information about whether a pair of communities has any members in common (intersection). We define community overlap features f_5 and f_6 as,

$$f_5 = \left|\phi_A(C_i) \cap \phi_A(C_j)\right|$$
$$f_6 = \left|\phi_B(C_i) \cap \phi_B(C_j)\right|$$

where $\left|\phi_k(C_i) \cap \phi_k(C_j)\right|$ is the number of individuals that are members of both C_i and C_j and are subscribers of carrier k. Community overlap features are expected to be useful in inferring unobserved communication, because nonzero values imply that there are individuals that belong to both

[1] Details of the community detection process, and the results of applying it to large-scale social networks, will be published elsewhere.

communities, increasing the likelihood of inter-community communication.

3) Cross-Community Call Frequency: The number of observed phone calls can be extracted from the CDR data. Let $\psi(a, b)$ be the number of phone calls observed between individuals a and b during the time period of observation. When working at the community level, we may also define $\Psi(C_i, C_j)$ to be the number of inter-community calls between members of communities C_i and C_j. Additionally, let $\Psi(\phi_k(C_i), \phi_k(C_j))$, with $k = A, B$ be the number of inter-community calls placed between community C_i –with all calls belonging to carrier A– and community C_j –with all calls coming from carrier B. For example,

$$\Psi(\phi_A(C_i), \phi_B(C_j)) = \sum_{\alpha=1}^{\left|\phi_A(C_i)\right|} \sum_{\beta=1}^{\left|\phi_B(C_j)\right|} \psi(C_{i,\alpha}, C_{j,\beta})$$

is the number of inter-community calls placed between communities C_i and C_j where all the calls from community C_i belong to carrier A and all the calls from community C_j belong to carrier B. We denote by $C_{i,\alpha}$ the individual of C_i with index α.

Hence, we create three additional features:

$$f_7 = \Psi(\phi_A(C_i), \phi_A(C_j)), f_8 = \Psi(\phi_A(C_i), \phi_B(C_j))$$
$$f_9 = \Psi(\phi_B(C_i), \phi_A(C_j))$$

Note that our goal is to accurately infer $y = \Psi(\phi_B(C_i), \phi_B(C_j))$, *i.e.,* all the phone calls between communities C_i and C_j belonging to the unobserved carrier B.

This set of cross-community features captures information about how frequently members of each carrier and each community call one another.

4) Cross-Community Call Duration: We define next a set of features that characterizes the total duration of calls between two communities, where we again distinguish the carrier for each caller. Let $\omega(a, b)$ be the total duration of calls between individuals a and b, and $\Omega(C_i, C_j)$ be the total duration of calls between communities C_i and C_j. Finally, let $\Omega(\phi_k(C_i), \phi_k(C_j))$ with $k = A, B$, be the total duration of calls between communities C_i and C_j and carriers A and B. For example,

$$\Omega(\phi_A(C_i), \phi_B(C_j)) = \sum_{\alpha=1}^{\left|\phi_A(C_i)\right|} \sum_{\beta=1}^{\left|\phi_B(C_j)\right|} \omega(C_{i,\alpha}, C_{j,\beta})$$

represents the total duration of inter-community calls between communities C_i and C_j, where all the calls from community C_i belong to carrier A and all the calls from community C_j are from subscribers of carrier B.

We can use this definition to define three additional features:

$$f_{10} = \Omega(\phi_A(C_i), \phi_A(C_j)), f_{11} = \Omega(\phi_A(C_i), \phi_B(C_j))$$
$$f_{12} = \Omega(\phi_B(C_i), \phi_A(C_j))$$

Index	Variables
1	Target variable y
2-5	The community size features
6-7	Community overlap features
8-10	Observable inter-community call frequency features
11-13	Observable inter-community call duration features
14-17	Observable intra-community call frequency features
18-21	Observable intra-community call duration features
22	Unobservable inter-community call frequency
23	Unobservable inter-community call duration
24-25	Unobservable intra-community call frequency
26-27	Unobservable intra-community call duration

TABLE I
VARIABLE INDICES FOR FIGURE 1

Call duration features have been shown to capture aspects of the strength of a social interaction not captured by interaction frequency[21]. Intuitively, if members of both communities tend to have long calls, unobserved members are more likely to have unobserved interactions.

Finally, we define a set of features that characterize each community individually.

5) Intra-Community Call Frequency: Given a pair of communities (C_i, C_j), we define a set of features that examines calls placed within each community, again for both carriers.

$$f_{13} = \Psi(\phi_A(C_i), \phi_A(C_i)), f_{14} = \Psi(\phi_A(C_i), \phi_B(C_i)),$$
$$f_{15} = \Psi(\phi_A(C_j), \phi_A(C_j)), f_{16} = \Psi(\phi_A(C_j), \phi_B(C_j))$$

This set of features captures information about how often members of the community place phone calls in general. Note that we do not define $\Psi(\phi_B(C_i), \phi_B(C_i))$ or $\Psi(\phi_B(C_j), \phi_B(C_j))$ as they are unobserved.

6) Intra-Community Call Duration: The last set of features measures the duration of calls placed within a community.

$$f_{17} = \Omega(\phi_A(C_i), \phi_A(C_i)), f_{18} = \Omega(\phi_A(C_i), \phi_B(C_i)),$$
$$f_{19} = \Omega(\phi_A(C_j), \phi_A(C_j)), f_{20} = \Omega(\phi_A(C_j), \phi_B(C_j))$$

This set of features characterizes how long members of each community speak for and the implied tie strengths. Again, we do not define $\Omega(\phi_B(C_i), \phi_B(C_i))$ or $\Omega(\phi_B(C_j), \phi_B(C_j))$.

7) Variable Correlation Matrix: Figure 1 displays the correlation matrix between the target variable y, and the 20 features presented here plus the 6 additional unobservable features as a heat map. There are a total of 27 variables, whose indices are summarized in Table I. The strongest correlations between the target variable y and the observable features are with feature f_7, which measures community overlap between subscribers of carrier B, and with f_9 and f_{10}, which measure inter-community call frequency where the two subscribers are from different carriers.

V. EXPERIMENTAL RESULTS

A. Experimental Design

Given the original data set of CDRs, we follow the procedure described in Section III-C to create the partially observed

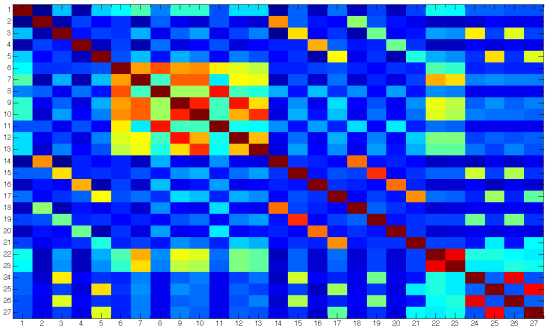

Fig. 1. [best seen in color] Correlations between variables described in Table I

data set used in our experiments. First, all calls in the 6-month period where both parties are subscribers of a single carrier (carrier A) are extracted, which yields a set of about 17 million phone calls. Next, 50% of the subscribers are randomly switched to a different carrier (carrier B) and all the calls between two subscribers of carrier B are considered to be hidden (about 4.3 million phone calls). This partially observed data set is used to automatically extract communities as explained in Section IV-B, yielding $6,963$ overlapping communities. The 20 features presented earlier are computed on the extracted communities, using only the observed data in the partially observed data set. Finally, we use the hidden data to compute the binary target variable y, which is non-zero when a pair of communities had at least one unobserved inter-community phone call during the 6-month period. About $24,000$ community pairs had inter-community communication. Each of these pairs served as one observation in the training data, where an observation consists of the target variable y and the 20 features associated with each community pair.

In order to infer the unobserved links, *i.e.,* community pairs that had unobserved inter-community communication, we use two different binary classifiers: an artificial neural network (ANN) and logistic regression. These two classifiers produce an estimation of the probability of unobserved links between two communities. Both yield very similar results, and in both cases, an improvement over a naïve majority-class classifier.

B. Artificial Neural Network

We use a three layer ANN with an input layer, a hidden unit layer (with 10 hidden units), and an output layer. The data is split into three subsets: 70% of the data is used for training the model, 15% for testing, and the remaining 15% for validation. All the features are used as input without applying any transformations. The neural network is then trained using scaled conjugate gradient backpropogation.

C. Logistic Regression

The binary classification problem of deciding whether or not a pair of communities (C_i, C_j) had some unobserved communication maps each pair of communities to the set $\{0, 1\}$. A logistic regression model can be used to perform binary classification by first fitting the data using the model and then defining a threshold value: probabilities below the threshold are assigned to the negative class (no unobserved communication) and probabilities equal to or above the threshold are assigned to the positive class (unobserved communication).

We fit a logistic regression model with all 20 features and use $p = 0.5$ as the threshold. The regression coefficients quantify the contribution of each feature to the predicted probability. Positive regression coefficients imply that larger values for the corresponding feature increase the probability that the community pair had unobserved communication.

The largest coefficients with our data set were found to be for the community overlap features. This result is intuitive, since members in communities that are largely overlapping seem to be more likely to interact with each other than members in communities without overlaps. The second largest coefficients corresponded to the call frequency features. Conversely, the smallest coefficients (which were near zero in value) were for call duration features.

D. Prediction Accuracy

Note that 87.47% of the community pairs in the training data do not have any communication. Therefore, we have a pronounced class-imbalance. In this case, the baseline naïve majority-class classifier would always predict that there was no communication between two randomly chosen communities. On the validation data, this strategy correctly predicts 87.47% of the time, as only 12.53% of the community pairs have some unobserved communication.

The accuracy of the predictions made by the ANN and the logistic regression model are marginally better than that of the naïve majority-class predictor: 89.9% for the artificial neural network and 89.84% for the logistic regression model with a threshold for the positive class at probability $p = 0.5$.

E. Lift

Next, we focus on the following task: can we identify all the community pairs that had some unobservable communication? A standard method for trying to find such pairs is to use the numeric output from the logistic regression or artificial neural network to rank the results in order of descending probability. The first N observations correspond to the N community pairs that were most likely to have had inter-community communication.

This methodology is widely used in marketing, for example when targeting the most likely candidates for a new marketing campaign[22]. The concept of *lift* [23] is commonly used in this setting to determine how much better than random a model performs when identifying the N most likely positive ($y = 1$) observations. Lift quantifies our ability to better identify a subset of true positives. The the lift factor is defined as the fraction of true positives in the validation set divided by the fraction of true positives in the training set.

Figure 2 depicts the lift factor for both the neural network and the logistic regression classifiers. The large plot shows the

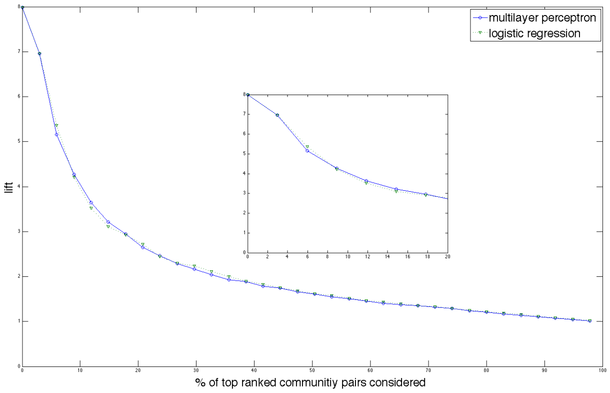

Fig. 2. Lift curve for the neural network and logistic regression classifiers.

lift factor for all the observations in the validation data set. The x-axis is the percentage of the ranked data considered according to the scores of both classifiers and the y-axis measures the lift factor for the first x % of the data. The small plot displays the first 20% of the data, which corresponds to the largest lift.

As the lift chart shows, a maximum lift of 8 is achieved for a small subset of the data. The lift then drops as more data points are included in the subset since the percentage of false positives rises. This is a typical example of increasing recall while decreasing precision. To the best of our knowledge, there are no other studies that infer unobserved links in a social network, either at the individual or community level, so we are not able to compare the results put forth here with any baseline other than random subsets.

VI. Discussion and future work

Unobserved social interaction is prevalent in a wide array of settings. In this paper, we have presented a new methodology for *inferring* unobserved links that already exist in a social network. We have also modeled interactions between *communities* rather than individuals and presented a method for recovering overlapping communities. At this higher level of abstraction and for a social network built from mobile phone calls, we have built two classifiers (neural networks and logistic regression) to determine which communities had unobserved intra-community communication. We have shown that both classifiers perform similarly, and are able to correctly identify a small subset of nodes where up to eight times more pairs had unobserved communication than a randomly chosen subset of the same size. For a larger subset consisting of 10% of all possible pairs, the classifiers correctly identify four times more pairs that communicated than a randomly chosen subset of the same size.

We consider the problem of recovering missing information about unobservable social interaction to be an important open issue. Identifying communities that have unobservable interactions can, for example, lead to better community identification and also allow for the discovery of social groups that share common interests.

References

[1] B. Wellman and S. D. Berkowits, *Social Structures: A Network Approach*. Cambridge University Press, 1998.

[2] 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*, vol. 12, no. 3, pp. 211–223, 2001.

[3] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[4] S. Gregory, "An algorithm to find overlapping community structure in networks," in *11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007)*. Springer-Verlag, 2007, pp. 91–102.

[5] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818.

[6] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *PROC.NATL.ACAD.SCI.USA*, vol. 101, pp. 2658–2663, 2004.

[7] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *PROC.NATL.ACAD.SCI.USA*, vol. 99, pp. 7821–7826, 2002.

[8] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, no. 6, p. 066111, Dec 2004.

[9] L. Getoor and C. Diehl, "Link mining: a survey," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 3–12, 2005.

[10] E. Zheleva, L. Getoor, J. Golbeck, and U. Kuter, "Using Friendship Ties and Family Circles for Link Prediction."

[11] A. Clauset, C. Moore, and M. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[12] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*. New York, NY, USA: ACM Press, 2003, pp. 556–559. [Online]. Available: http://dx.doi.org/10.1145/956863.956972

[13] M. Lahiri and T. Berger-Wolf, "Structure prediction in temporal networks using frequent subgraphs," in *IEEE Symposium on Computational Intelligence and Data Mining (CIDM 07)*, 2007.

[14] J. O'Madadhain, J. Hutchins, and P. Smyth, "Prediction and ranking algorithms for event-based network data," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 23–30, 2005.

[15] A. Popescul and L. Ungar, "Statistical relational learning for link prediction," in *IJCAI workshop on learning statistical models from relational data*, 2003.

[16] D. Kempe, J. Kleinberg, and A. Kumar, "Connectivity and inference problems for temporal networks," *Journal of Computer and System Sciences*, vol. 64, no. 4, pp. 820–842, 2002.

[17] M. Cebrián and E. Frías-Martínez, "Word-of-Mouth Algorithms: What you don't know will hurt you," *Telefonica Research Technical Report*, 2008.

[18] M. Lahiri, A. Maiya, R. Sulo, and T. Habiba, "The Impact of Structural Changes on Predictions of Diffusion in Networks," in *IEEE International Conference on Data Mining Workshops, 2008. ICDMW'08*, 2008, pp. 939–948.

[19] F. Provost, "Machine learning from imbalanced data sets 101," *AAAI Technical Report WS-00-05*, 2001.

[20] C. Patrick, "Dataquest insight: Social network analysis is proving a useful tool for telecommunications operators," *Gartner Inc.*, 2009.

[21] M. S. Granovetter, "The strength of weak ties," *The American Journal of Sociology*, Jan 1973.

[22] G. Piatetsky-Shapiro and B. Masand, "Estimating campaign benefits and modeling lift," in *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 1999, pp. 185–193.

[23] G. Piatetsky-Shapiro and S. Steingold, "Measuring lift quality in database marketing," *SIGKDD Explor. Newsl.*, vol. 2, no. 2, pp. 76–80, 2000.