Social Network Discovery based on Sensitivity Analysis

Tarik Crnovrsanin, Carlos D. Correa and Kwan-Liu Ma
Department of Computer Science
University of California, Davis
tecrnovrsanin@ucdavis.edu, {correac,ma}@cs.ucdavis.edu

***Abstract*—This paper presents a novel methodology for social network discovery based on the sensitivity coefficients of importance metrics, namely the Markov centrality of a node, a metric based on random walks. Analogous to node importance, which ranks the important nodes in a social network, the sensitivity analysis of this metric provides a ranking of the relationships between nodes. The sensitivity parameter of the importance of a node with respect to another measures the direct or indirect impact of a node. We show that these relationships help discover hidden links between nodes and highlight meaningful links between seemingly disparate sub-networks in a social structure. We introduce the notion of implicit links, which represent an indirect relationship between nodes not connected by edges, which represent hidden connections in complex networks. We demonstrate our methodology on two social network data sets and use sensitivity-guided visualizations to highlight our findings. Our results show that this analytic tool, when coupled with visualization, is an effective mechanism for discovering social networks.**

***Keywords*-Social networks, Sensitivity analysis, Markov centrality, Network visualization.**

I. Introduction

Social networks have emerged as one of the most popular applications on the Web, as capitalized by the popularity of sites such as Facebook and Flicker. Their analysis and visualization, however, has only become more complex over time. When coupled with node-link visualizations, the analysis of social networks is a powerful technique for understanding complex social structures. However, node-link diagrams, often derived from extracted data such as calls and common interests, do not always represent the underlying social structure that exists in the real world. Even when two nodes do not appear to be linked in the visualization, they may carry influence on each other in the real social network.

In an attempt to rank nodes in a social network beyond the information that direct links offer, people have applied the concept of importance or centrality. Although some have proposed simple metrics for centrality based on degree and betweenness, global measures (e.g., PageRank and Markov centrality) prove to be robust estimators of importance in a social network. However, there is a question that arises when considering importance: how much impact has any given node on the importance of another? Answering this question, which we dubbed *importance sensitivity*, is the focus of this paper. We present a method for deriving the sensitivity parameters of a social network based on Markov importance [1]. Analogous to Markov importance, which ranks the nodes in a social network, the sensitivity parameters rank each pair of nodes. In this way, we can extract and visualize the most important node-to-node relationships that may or may not be represented explicitly in the social network graph. In particular, we are interested in extracting *implicit links* between nodes, which are links between two nodes that are not represented as an edge, but that exhibit a large sensitivity. Implicit links may indicate indirect relationships between centric nodes in a social network in terms of the sub-networks they share. In other cases, implicit links show bridging nodes between seemingly disconnected sub-networks. Through a couple of examples, we show that these newly found links are important for social network discovery.

To validate our approach, we applied the methodology to two data sets. One is a synthetic social network created for the VAST challenge 2008 [2], which contains a hidden social network that must be found through analytic and visual means. We show that our approach gives the answer by discovering the hidden relationships among the actors of interest. The second data set is the MIT reality network [3], consisting of communication, proximity and activity information from 100 subjects at MIT. Our approach simplifies the analysis of the network by discovering the most important links between seemingly disparate sub-networks.

II. Related work

The analysis of social networks has a long and exhaustive treatment in the literature, from the statistical [4], [5] and visual perspectives [6], with wide applications in the biological sciences, sociology and information systems [7].

A vast number of statistical properties for measuring social networks has been proposed, including clustering, degree distributions and centrality [4]. Centrality determines the relative importance of a node in a network. Some of these metrics are degree, betweenness and closeness centrality [8]. A more sophisticated approach is eigenvector centrality, used widely in Web ranking, as described in the seminal papers detailing the PageRank [9] and HITS algorithms [10]. White and Smyth describe an alternative approach, called Markov centrality, which considers a social network as a Markov chain [1], based on the mean first-passage time metric

[11]. These methods are inherently global and help extract important nodes of an otherwise vast network. In this paper we derive a sensitivity analysis of the Markov centrality metric. Similar to node centrality, sensitivity coefficients provide an idea of the importance of links in a network. Unlike the node-centered metrics, these coefficients rank the node-to-node connections, even when no explicit link exists between two nodes.

Other alternatives include link prediction for time-varying networks [12]. Sensitivity coefficients answer to the hypothetical question of how would centrality change if there is a small change in a node's set of connections. Given a time-varying network, this could be derived via statistical prediction. In the absence of temporal networks, an analytic derivation (or model fitting) is necessary. In this paper, we opt to derive an analytic expression to compute the sensitivity parameters. In the context of protein networks, Goldberg and Roth use mutual clustering to rank the links between nodes [13]. Our approach, although focused on centrality, could be applied to other similar metrics.

A recent effort for analyzing social networks uses visual metaphors [14], [15]. Dwyer et al. compare the different centrality measures using visual means [16]. van Ham and Wattenberg, on the other hand, exploit centrality to guide the visualization of small world graphs [17]. In this paper, we guide the visualization of node links using the sensitivity coefficients. We show that, depending on these values, we can tag derived implicit links between otherwise disconnected nodes or highlight existing connections.

III. Methodology

Our approach to finding hidden relationships consists of two stages. First, an analysis stage computes the Markov importance of each node and their sensitivity parameters with respect to the degree of each node. Second, we encode visually each node-pair depending on the magnitude of the sensitivity parameter and whether there exists an edge for that pair.

A. Markov Importance

As described in [1], the importance of a node in a social network graph can be measured as the mean first-passage time in the social network graph when understood as a Markov chain. The mean first passage time can be defined as the expected number of nodes a message starting from a given node s encounters until it reaches another node t for the first time [18]. White and Smyth found that this mean first passage can be computed as a matrix:

$$M = (I - Z + EZ_{dg})D \tag{1}$$

where I is the identity matrix, E is a matrix containing all ones, D is a diagonal matrix where each element in the diagonal is the reciprocal of the stationary distribution $\pi(v)$ of a node v, and Z is the so called fundamental matrix, given by:

$$Z \quad = \quad (I - A - e\pi^{\mathsf{T}})^{-1} \tag{2}$$

where A is the network adjacency matrix and π is a column vector of the stationary probabilities. The importance of a node v is the inverse of the average of the corresponding column in M:

$$I(v) = \frac{n}{\sum_{s \in V} m_{sv}} \tag{3}$$

B. Sensitivity Parameters

To find hidden relationships between nodes, we look at the sensitivity parameters of the importance metric of one node with respect to the degree of the other. In this way, we can measure how much would the importance change if we were to add or remove an edge to another node. Important hidden relationships arise when the sensitivity parameters are larger than a given threshold. According to sensitivity analysis methods, one can describe the sensitivity of a given function with respect to another variable as its partial derivative:

$$s_{ij} = \frac{\partial I_i}{\partial \Lambda_j} \tag{4}$$

where I_i is the importance of node i and Λ_j is the degree of node j.

The sensitivity coefficients of the entire matrix M with respect to a node i can be computed as :

$$\frac{\partial M}{\partial \Lambda_i} = (I - Z + EZ_{dg})\frac{\partial D}{\partial \Lambda_i} + (-\frac{\partial Z}{\partial \Lambda_i} + E\frac{\partial Z_{dg}}{\partial \Lambda_i})D \tag{5}$$

The derivative of the fundamental matrix is computed as:

$$\frac{\partial Z}{\partial \Lambda_i} = -Z(\frac{\partial Q}{\partial \Lambda_i} - e\frac{\partial \pi}{\partial \Lambda_i}^{\mathsf{T}})Z \tag{6}$$

where $\frac{\partial Q}{\partial \Lambda_i} = -\frac{\partial A}{\partial \Lambda_i}$ is the derivative of the probability matrix $Q = I - A$, computed via finite differences, and $\frac{\partial \pi}{\partial \Lambda_i}$ is the partial derivative of the stationary probabilities with respect to the degree of node i. Since $Q\pi = 0$, differentiating at both sides yields

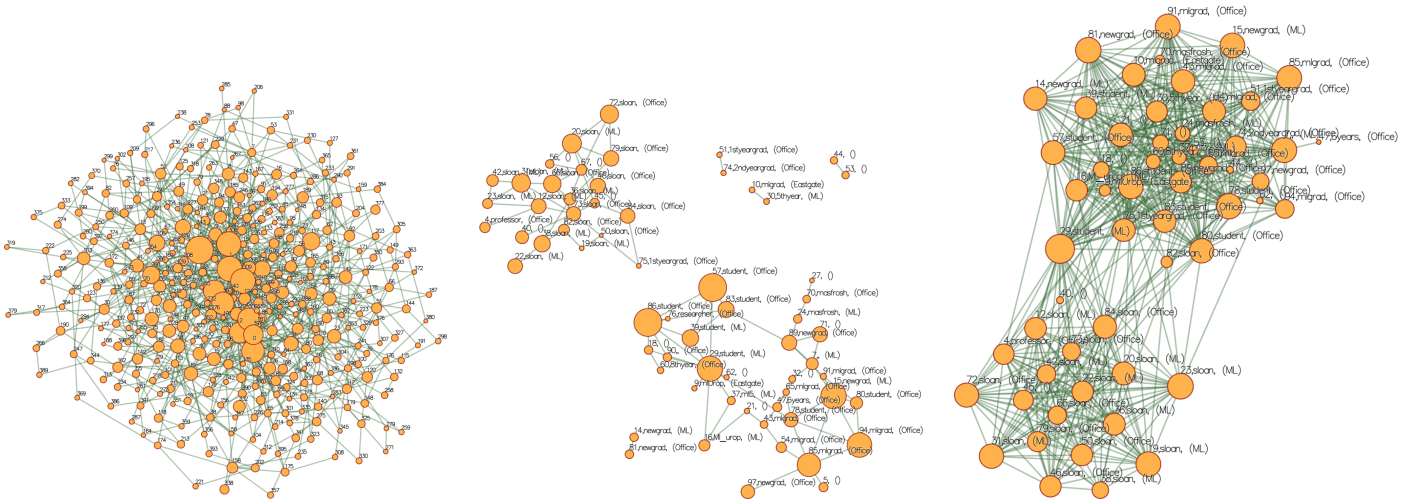
$$Q\frac{\partial \pi}{\partial \Lambda_i} + \frac{\partial Q}{\partial \Lambda_i}\pi = 0 \tag{7}$$

from which we can obtain the derivative of Q, as described by Haverkort [19]:

$$\frac{\partial \pi}{\partial \Lambda_i} = -Q^{-1}\frac{\partial Q}{\partial \Lambda_i}\pi \tag{8}$$

The derivatives of A can be approximated via finite differences: $\partial A/\partial \Lambda_i \approx A'_i - A_i$, where A'_i is the matrix of probabilities that results when adding 1 to the degree of node i (Both A' and A are normalized to represent a matrix of probabilities).

The matrix of coefficients, formed by the partial derivatives of M_i with respect to the degree of each node,



(a) VAST Challenge (b) MIT Call data (c) MIT Proximity data

Figure 1. Data sets used for validation. (a) VAST Challenge data set consisting of 400 nodes. The highly connected network makes visual-based analysis difficult (b) MIT communication data consisting of 100 nodes. We can see two big disconnected sub-networks. (c) MIT proximity data for one week. In terms of proximity, some of the nodes in (b) now appear to be linked together, but the high density of links destroys the readability of the network.

represents a ranking of the edges in terms of impact of changes in a node to the importance of another. The most important edges are those where the absolute value of the sensitivity coefficient is largest. Note that the derivatives can also be negative, meaning a negative impact of a node into another.

C. Visualization

Once we have found the sensitivity parameters, we can guide our visualization to help us discover hidden relationships. We can distinguish three types of edges. Let the user define a sensitivity threshold τ. Let us denote σ_{ij} as the sensitivity of a node i with respect to another node j.

- When $\sigma_{ij} \geq \tau$ and $ij \in E$, where E is the set of edges in the social network, the sensitivity indicates an important connection between nodes that are explicitly linked. These are shown in the figures as thick blue lines.
- When $\sigma_{ij} \geq \tau$ and $ij \notin E$, the sensitivity indicates an important connection that is *implicitly* represented in the network. These are hidden relationships which may provide additional insight into the social structure. These are shown in the figures as dashed lines.
- When $\sigma_{ij} < \tau$, the sensitivity does not indicate an important connection and it is not shown to avoid clutter.

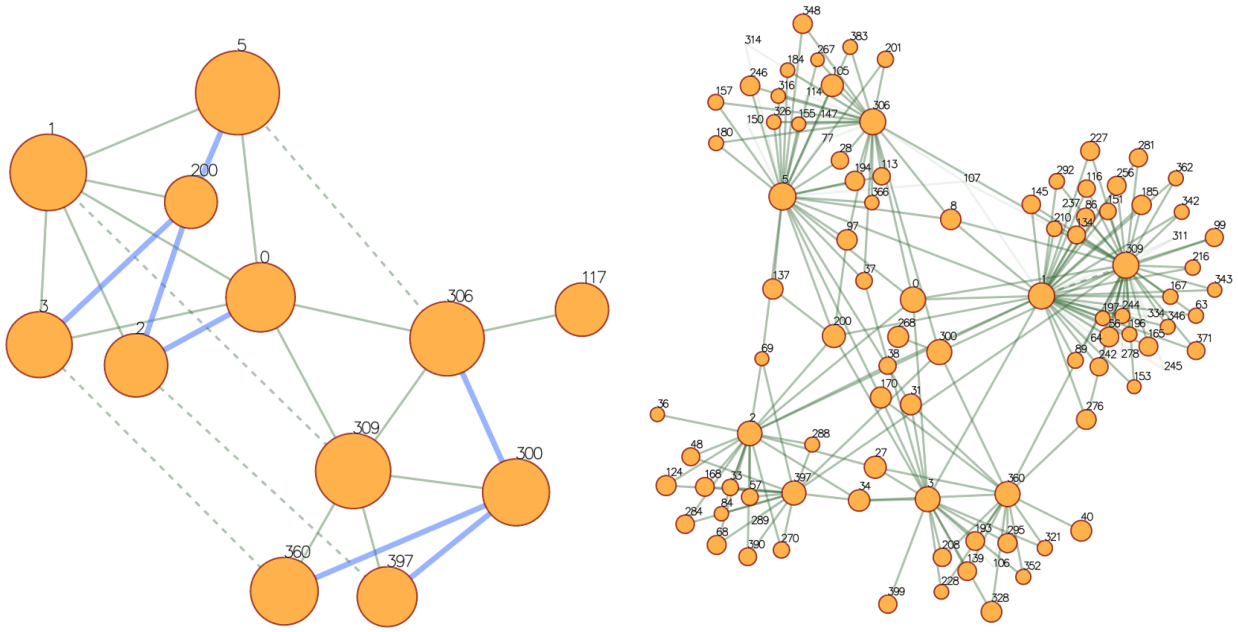
IV. Results

To validate our results, we present two case studies using the VAST challenge social network data set and the MIT Reality data set. The objective was to see if any hidden relationships could be found with the sensitivity analysis.

A. VAST Challenge

The VAST Challenge social network data set consists of communication logs among 400 unique cell phones during a span of a 10 day period in June 2006. The social network is a synthetic data set detailing the communication patterns of a fictitious organization centered around a person named Ferdinando Catalano, associated in the data set to the device with identifier 200. The purpose of the challenge task was to characterize his social network.

The entire social structure is dense, and each node shares at least two connections with any other node. This makes visual analysis and discovery a tedious task. See for example the node-link diagram in Figure 1(a). The clutter makes the structure illegible. As an initial step, we computed the Markov importance on the network and threshold it to show us the highest ranked nodes, leading to the visualization in Figure 2(a). Now we can see the immediate social network around identifier 200, consisting of nodes $1, 2, 3$ and 5. According to this metric, another sub-network appears (nodes $300, 306, 309, 360$ and 397) connected to the previous sub-network via node 0. To further investigate the relationship between these two sub-networks, we turn to sensitivity analysis. Our approach generates a ranking of the edges as depicted in Figure 2(a). Dashed lines show a strong connection between nodes that were not explicitly linked via calls. This shows a mutual impact between the importance of the two connected nodes. For instance, a change in node 1 affects the importance of node 309. A similar relationship was discovered for the node pairs $5 - 306$, $2 - 397$ and $3 - 360$. In addition, the blue thick lines highlight existing links that are also important. In this case, we noticed a symmetry between the groups formed by $2 - 3 - 5 - 200$



(a) (a) (b) (b)

Figure 2. (a) Discovered links in the core network of the VAST challenge. The dashed lines show a strong connection between certain pairs of nodes that are not otherwise explicitly connected. Note also the mirroring of the blue links between the groups 200,3,2,5 and 300,397,360,306. (b) Validation of the VAST Challenge. Further inspection of the data set shows that certain pairs of nodes (e.g., 1 and 309) share a sub-network. This situation was identified as the fact that the same people use two different cell phones to communicate. The pairing of cell phones per person, inferred from the nodes at the center of the common sub-networks, are precisely the one detected in Fig.2(a) (1-309, 5-306, 2-397 and 3-360)

and $360 - 397 - 306 - 300$.

To validate these findings, we plotted the sub-network formed by the two-degrees of separation from identifiers 200 and 300. The result, as shown in Figure 2(b), shows the overall social network sought after in the challenge. Notice that the four pairings that we found with our approach appear as the loci of four sub-networks. For example, node 5 and 306 share most of the connections among themselves. This suggests that the pairings correspond to device identifiers belonging to the same people. With our approach, these links become evident as depicted in Figure 2(a). Our results are further validated with respect to the answers obtained by other challenge participants [20].

This case study shows a good example of the uses of our approach for social network discovery. Node importance metrics by themselves cannot fully explain the relationships between two sub-networks of interest. In this case, Markov centrality highlights the most central nodes in the social structure, but cannot convey the *implicit* relationships that arise from each of their own sub-networks. Looking at the sensitivity coefficients, we are able to extract that information.

B. MIT Reality

The MIT reality data set keeps track of 100 subjects at MIT over the course of the 2004-2005 academic year, containing over $350,000$ hours of human activity. Although seemingly small in terms of communication (64 users, as shown in Figure 1(b)), the data set is rich with information about proximity, infrastructure and activity logs. One of the key questions is to be able to characterize the social structure and find out if the topology can be inferred from proximity data alone.

Figure 1(b) shows the node-link diagram for the communication network. Two different networks appear in the graph, corresponding to people in the MIT Media Lab (bottom right) and the MIT Sloan business school (top left), respectively. Looking at calls alone, we see that there is no apparent connection between these two sub-networks. When looking at proximity data, as shown in Figure 1(c), we see a much more connected graph, given that the two buildings are adjacent. Even when considering a single week, the network is cluttered and difficult to read. The purpose of our study was to discover important links between the two sub-networks based on proximity alone. We used sensitivity on the proximity data to rank the connections between nodes in the call network. The call network provides us a more compact view of the social structure, while proximity provides an aggregated connection between disjoint sub-networks. The result is depicted in Figure 3(a). When considering only the highest ranked links, we see a majority of connections within each of the sub-networks. This expected behavior confirms the mirroring of the calls in the proximity data set. As an

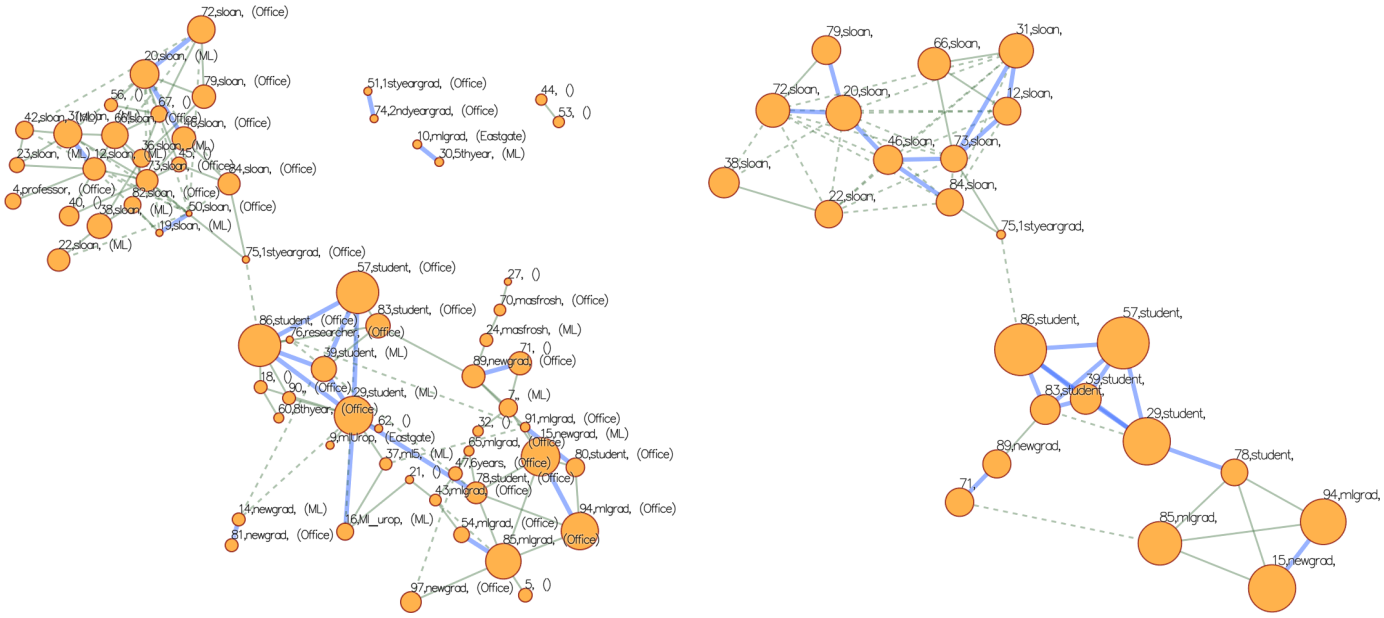


Figure 3. (a) Result of highlighting important edges on the MIT reality data set. Although a number of links are highlighted, we immediately discover the connection between the two sub-networks, as shown between nodes 75 and 86. They indicate a commonality between the two networks that is difficult to see from the proximity data. (b) Combining centrality-based visualization with sensitivity parameters. By thresholding the network with respect to centrality, we get a more compact representation of the social structure. The sensitivity parameters highlight those relationships that are more important.

interesting result, we see a dashed line between nodes 75 and 86 that connects the two sub-networks. Node 75 can be identified as the only graduate student in the Sloan sub-network. Their activity patterns seem to overlap with those of other students in the Media Lab sub-network. However, we cannot tell with certainty that these two nodes are in close proximity of each other, but rather that they have a strong impact on the importance of each other. When considering the social structure of two adjacent but different buildings, it is likely that the students will provide the bridge between the two. A more compact diagram, obtained by showing only the nodes with highest ranking, is depicted in Figure 3(b). The simplified representation also shows the tight connectivity in the Sloan sub-network (everyone impacts almost each other), reflected in the number of implicit links (dashed lines), compared to the Media Lab sub-network. Sensitivity coefficients are therefore a metric for analyzing the asymmetry of social networks.

We also noticed strong links between connected nodes in the call graph. These are depicted as thick blue lines. For example, we notice a clique formed by students $29, 39, 57$ and 86. The sensitivity analysis therefore provides a starting point for further discovery.

V. Discussion

Our approach provides important information about hidden relationships within social networks. Just like importance metrics rank nodes, our method ranks the links between nodes, even those that are not explicitly represented. In this paper, we focused on the visualization of *implicit links* (dashed lines) as a means for social network discovery. The emphasis on important explicit links (thick blue lines) steered our attention towards cliques or sub-networks of interest. There are many aspects of these sensitivity parameters that can be further explored. Unlike the importance function, sensitivity parameters can be negative. A change of a node can have a detrimental impact on the importance of another, say, by becoming more important and out-ranking the other node. In our examples, we have only considered the positive impact between two nodes as a measure of relative importance. However, some applications may benefit from the signed nature of the sensitivity parameters. For example, the distribution of the sign of the sensitivity parameters may give hints about the asymmetry of the network. In addition, sensitivity metrics can lead to an uncertainty analysis of a social network. Although social networks are commonly built on actual communication, proximity or activity patterns, they rely on assumptions about social interaction that may not be captured in the form of digital data. Therefore, the analysis of the mutual impact of any two given nodes provides a tool to quantify the uncertainty of network metrics, such as centrality.

One of the limitations of our approach is the reliance on a global algorithm to compute importance and its derivatives. Markov centrality requires a computational cost of $O||V||^3)$, where $||V||$ is the number of nodes in the network. This makes this approach unfeasible for extreme scale social

networks. As an alternative, one can consider the sensitivity coefficients among clusters and groups instead of individual nodes or apply the model locally for partial analysis.

VI. Conclusion

We have presented a novel methodology for discovering relationships in social network graphs. Based on the sensitivity coefficients of the Markov importance of a node, our approach identifies those edges with the largest mutual impact in importance. This impact can indicate a number of things in the social network, such as presence of hidden connections between seemingly separate sub networks. One of the applications we explored was the guidance of visualization to aid in discovery. However, our approach can be extended as a means to filter out unimportant parts of the data and focus on connections that may not be extracted using traditional filtering methods. Furthermore, as demonstrated in the MIT reality data set, we can use sensitivity coefficients from one data set to complement the view of another, such as proximity, communication or activity. The exploration of sensitivity parameters of commonly used analytical metrics leads to novel ways of looking at social network data, and offers insight for both local (e.g., the relationship between two particular nodes) and global inquiries (e.g., the overall social structure). Coupled with visualization, these analytic tools help understand the complex relationships between actors in a social network.

Acknowledgment

This research was supported in part by the U.S. National Science Foundation through grants CCF 0808896 and CCF-0811422 and by Hewlett-Packard Laboratories. The authors would like to thank the VAST 2008 challenge chairs and the MIT reality mining team for their data sets.

References

[1] S. White and P. Smyth, "Algorithms for estimating relative importance in networks," *International Conference on Knowledge Discovery and Data Mining*, vol. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, no. 2, pp. 266 – 275, 2003.

[2] http://www.cs.umd.edu/hcil/VASTchallenge08/, "IEEE VAST 2008 challenge," 2008.

[3] http://reality.media.mit.edu, "MIT reality mining," 2008.

[4] M. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167 – 256, 2003.

[5] S. Wasserman, K. Faust, and M. Granovetter, *Social Network Analysis: Methods and Applications*. Cambridge Univ. Press, 1994.

[6] L. Freeman, "Visualizing social networks," *Journal of Social Structure*, vol. 1, no. 1, 2000.

[7] J. P. Scott, *Social Network Analysis: A Handbook*. SAGE Publications, January 2000.

[8] L. C. Freeman, "Centrality in social networks: Conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215 –239, 1979.

[9] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1, pp. 1–7,107–117, April 1998.

[10] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[11] J. Kemeny and J. Snell, *Finite Markov Chains*. Springer Verlag, 1976.

[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*, 2003, pp. 556–559.

[13] D. S. Goldberg and F. P. Roth, "Assessing experimentally derived interactions in a small world." *Proc Natl Acad Sci U S A*, vol. 100, no. 8, pp. 4372–4376, April 2003.

[14] M. Garland and G. Kumar, "Visual exploration of complex time-varying graphs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 805–812, 2006.

[15] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*. Upper Saddle River, NJ: Prentice Hall PTR, 1998.

[16] T. Dwyer, S.-H. Hong, D. Koschützki, F. Schreiber, and K. Xu, "Visual analysis of network centralities," in *APVis '06: Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation*. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2006, pp. 189–197.

[17] F. van Ham and M. Wattenberg, "Centrality based visualization of small world graphs." *Comput. Graph. Forum*, vol. 27, no. 3, pp. 975–982, 2008.

[18] U. Brandes and T. Erlebach, *Network Analysis : Methodological Foundations (Lecture Notes in Computer Science)*. Springer, March 2005.

[19] B. Haverkort and A. Meeuwissen, "Sensitivity and uncertainty analysis of markov-reward models," *IEEE Transactions on Reliability*, vol. 44, no. 1, pp. 147–154, Mar 1995.

[20] G. Grinstein, C. Plaisant, S. Laskowski, T. O'Connell, J. Scholtz, and M. Whiting, "VAST 2008 challenge: Introducing mini-challenges," *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, pp. 195–196, Oct. 2008.