egoStellar: Visual Analysis of Anomalous Communication Behaviors from Egocentric Perspective

Mei Han[1(✉)], Qing Wang[1(✉)], Lirui Wei[1(✉)], Yuwei Zhang[1(✉)], Yunbo Cao[2(✉)], and Jiansu Pu[1(✉)]

[1] Visual Analytics of Big Data Lab, UESTC, Chengdu, China
hanmeil993@126.com, vincent.w.qing@gmail.com, 413095866@qq.com, 494479914@qq.com, jiansu.pu@uestc.edu.cn

[2] School of Information and Software Engineering, UESTC, Chengdu, China
uestc2008@126.com

Abstract. Detection and analysis of anomalous communication behaviors in cellar networks are extremely important in identifying potential advertising agency or fraud users. Visual analytics benefits domain experts in this problem for its intuitiveness and friendly interactive interface in presenting and exploring large volumes of data. In this paper, we propose a visual analytics system, egoStellar, to interactively explore the communication behaviors of mobile users from an ego network perspective. Ego network is composed of a centered individual and the relationships between the ego and his/her direct contacts (alters). Based on the graph model, egoStellar presents an overall statistical view to explore the distribution of mobile users for behavior inspection, a group view to classify the users and extract features for anomalous detection and comparison, and a ego-centric view to show the interactions between an ego and the alters in details. Our system can help analysts to interactively explore the communication patterns of mobile users from egocentric perspectives. Thus, this system makes it easier for the government or operators to visually inspect the massive communication behaviors in a intuitive way to detect and analyze anomalous users. Furthermore, our design can provide the researchers a good opportunity to observe the personal communication patterns to uncover new knowledge about human social interactions. Our proposed design can be applied to other fields where network structure exists. We evaluated egoStellar with real datasets containing the anomalous users with extremely large contacts in a short time period. The results show our system is effective in identifying anomalous communication behaviors, and its front-end interactive visualizations are intuitive and useful for analysts to discover insights in data.

Keywords: Mobile users · Ego-centric · Anomalous users · Visual analysis

C.-Y. Chang et al. (Eds.): ICS 2018, CCIS 1013, pp. 280–290, 2019.
https://doi.org/10.1007/978-981-13-9190-3_29

1 Introduction

The booming of information and communication technology nurtures the big data era [1]. Among all these large volumes of data, communication data generated from celluar network recording how people interact with each other though mobile phones. The accumulation of such mobile communication records introduces new mechanisms for experts to study the human communication behaviors. And analyzing these behaviors not only helps finding the common communication patterns for mobile users, but more importantly facilitates the detection and analysis of customers with anomalous behaviors in communication networks who are potential advertising agencies or fraud users. Visual analytics benefits domain experts in this problem for its intuitiveness representation of context information and additional evidence via interactive interface for result analyzing and exploring. The Ego Networks (ENs) examine the ties connecting a target individual (ego) and his/her direct contacts (alters). Most of the extant research on communication networks are from the overall perspective regardless of the personal network features, and they were carried out based on either statistics or machine learning methods [5–7]. In this paper, we propose egoStellar to explore the communication behaviors of mobile users from an ego network perspective. Specifically, we extract the ECNs from the communication data, and portray the ECNs with six network metrics [11]. In order to explore anomalous behaviors via the visual inspection of ego-centric networks, we further design three views for interactive investigations: the first view is the statistical view, which uses the interactive scatter design to capture the holistic correlations and distributions of different ECN features for all egos. The second view is the group view, which uses pixel-based design to classify different ECNs into groups. Last but not least, we propose the third ego-centric view for uses, which shows the proportions of local and alien, unidirectional and bidirectional alters together with the interactions between the bidirectional alters by applying a new novel glyph-based design shaped from galaxy. In summary, our contributions in this paper are: (1) **System:** We build ECNs based on the communication data and introduce a novel visual analytics system for interactively exploring the mobile users' communication behaviors from ego network perspectives. (2) **Visualization:** We propose a new novel glyph design and layout algorithms shaped from galaxy to efficiently detect and analyze, and compare users with different communication patterns. Our design helps the experts to grasp the overall, the group-level, and the personal level of ECN status of the users thus facilitates the anonymous users detection and analysis. (3) **Evaluation:** We have evaluated egoStellar with real datasets containing the anomalous users with extremely large contacts in a short time period to demonstrated its effectiveness and usability. Through quantitative measurements of the design performance and qualitative interviews with domain experts, the results show our system is effective in identifying anomalous communication behaviors, and its front-end interactive visualizations are intuitive and useful for analysts to discover insights in data.

2 Related Work

The widespread of mobile communication and Online Social Network (OSN) accumulates the relevant data so that we are able to study the social networks at large scale [5, 6]. Onnela et al. [12] uncovered the existence of the weak tie effect. Eagle et al. [6] found it possible to infer 95% of friendships accurately based only on the mobile communication data. Saramäki et al. [13] showed that individuals have robust and distinctive social signatures persisting overtime. Wang et al. [11] studied the communication network from egocentric perspective, and find that the out-degree of a user plays a crucial role in affecting its ECN structure. As illustrated above, much attentions have been paid to uncovering the overall features of the ego communications networks. Ego network has also been a heated topic in the information visualization community recently. Shi et al. [14] proposed a new 1.5D visualization design to reduce the visual complexity of dynamic networks without sacrificing the topological and temporal context central to the focused ego. Liu et al. [15] raised a constrained graph layout algorithm "EgoNetCloud" on dynamic networks to prune, compress, and filter the networks in order to reveal the salient part of the network. Wu et al. [16] presented a visual analytics system named "egoSlider" for exploring and comparing dynamic citation networks from macroscope, mesoscope, and microscope levels. Cao et al. [17] proposed "TargetVue", which detected the anomalous users of online communication system via unsupervised learning. Liu et al. [18] introduced "egoComp", the storyflow-like links design, into the node-link graph in order to reveal the relations between two ego networks. Most of the research mainly focus on visualizing the statistical features of the peer to peer interactions within the OSNs or citation networks, but the research on detailed ego communication networks and communication patterns are still insufficient.

3 Data and Methods

He call detail records are collected by mobile operators for billing and network traffic monitoring. The basic information of such data sets contains the IDs of callers and callees, time stamps, call durations, base station numbers, charge information and so on. The dataset used in this study covers 7 million people of a Chinese provincial capital city for half a year spanning from Jan. to Jun. 2014. According to the operator the users choose, all of the users can be divided into two categories, namely, the local users (customers of the mobile operator who provide this data set) and the alien users (customers from the other operators). The mobile communication network can be modeled as a directed graph G(V; E) with the number of nodes and links being $|V| = N$ and $|E| = L$, respectively. Link weight is defined as wij for a directed link lij, which is the number of calls that user i has made to user j, and it represents the link strength between two users.

People usually make calls to maintain their social relationships [13]. The directions of communication divided the alters into two sets for an ego *i*: the in-contact set C_i^{in} and the out-contact set C_i^{out}. The size of C_i^{in} and C_i^{out} are in-degree k_i^{in} and out-degree k_i^{out}, respectively. k_i^{in} represents the ECN size ego i maintains, and k_i^{in} can reflect the

influence of ego i in the network. In this paper, we mainly focus on k_i^{out}, because it represents the number of alters an ego intends to spend cognitive resources to maintain. We further define the node weight of an ego as $W_i^d = \sum_{j \in C_i^{out}} w_{ij}^d$ to indicate the total amount of energy an ego spend on maintaining his/her social relationships. In fact, the call durations are also important in communication behaviors and the link weight in "duration" perspective can be defined as $W_i = \sum_{j \in C_i^{out}} w_{ij}$, where w_{ij}^d is the call duration from i to j. To further investigate the properties of the ECNs, another three metrics are also introduced, namely, average node weight $\overline{w}$, attractiveness balance η, and tie balance θ. For ego i, the average node weight $\overline{w_i}$ is defined as:

$$\overline{w_i} = \frac{1}{k_i^{out}} \sum_{j \in C_i^{out}} w_{ij} \tag{1}$$

where w_{ij} is the weight of link l_{ij}, and k_i^{out} is the size of ECN. This metric indicates the average emotional closeness between an ego and the alters [13, 19]. Considering the communication directions, we introduce the attractiveness balance (AB) to measure such relationships between an ego and the network. It is defined in a straight forward way:

$$\eta_i = \frac{k_i^{in}}{k_i^{out}} \tag{2}$$

The attractiveness balance $\eta = 1$ means that the number of contacts a user calls is equal to the number of contacts who call him/her, suggesting the balance of the attractiveness. Large η implies strong attractiveness of an ego while small η refers to a weaker attractiveness. Apart from the attractiveness balance, communication directions also distinguishes bidirectional alters (who appear in both C_i^{in} and C_i^{out} from the unidirectional ones (who only appear in either C_i^{in} or C_i^{out}). Usually, the reciprocal relationships are stronger than the unidirectional relationships, thus they can be viewed as strong and weak ties [21]. Thus, we introduce another structural balance metric named tie balance (TB), which is defined as the Jaccard distance[22] between C_i^{in} and C_i^{out}. Mathematically, it reads:

$$\theta_i = \frac{\left| C_i^{in} \cap C_i^{out} \right|}{\left| C_i^{in} \cup C_i^{out} \right|} \tag{3}$$

$\theta = 1$ means all of ego i's direct contacts have bidirectional links with ego i, while $\theta = 0$ means ego i even has no reciprocal contacts. Of course, the above two kinds of ECNs are all extremely imbalance.

4 Visual Analytics System

4.1 System Overview

In this section, we introduce "egoStellar", whose design borrows the idea of galaxy. Like our solar system, an egocentric network is composed of a centered ego (like the sun), and all the other alters around him/her (like the other planets). The design goal of this visual analytics system is to give the analysts different levels of mobile users' calling behaviors: from the overall level of statistics, via group level statistics, to egocentric communication behaviors. To achieve this goal, we design 3 views for the corresponding level. Figure 2 illustrates the system architecture and the data processing pipeline of this visual analytics system. The system has two main parts as illustrated in Fig. 2(a), and they are "Computing End" and "Visual Representation End", which are connected via network. "Computing End" is a parallel computing cluster, which is composed of a Hadoop Distributed File System (HDFS), a customized Apache Spark parallel computing platform [24], and a "CompAgent". "CompAgent" receives computing tasks from the "Visual Representation", and cache the intermediate computing results for it. The "VisAgent" receives the data, and transmit the computing tasks to the "CompAgent". It can also transform data for visualization and send them to the "User Interface".

4.2 Visual Design

In this section, the three views are described concecutively. Firstly, the Statistical View is present to show the distribution of users according to their egonetwork size, which is from the macroscopic perspective; Secondly, the users are divided into different groups according to the correlation between the ego network size and communication frequency, which is the mesoscopic perspective; Most importantly, the behaviors of the specific users are shown in the Egocentric View, which is the microscopic perspective.

Statistical View. Due to the scaling problem, it is not easy to observe the rare items in the traditional distribution chart, and log scale suffers from its unintuitiveness. Rare items are significant for detecting abnormal users in our case, thus we design the multi-scale distribution view, in which we show the distribution for the majority of the population, and use bubbles to represent the rare items. This Statistical View is more efficient and practical than directly visualizing such communication data for quickly grasping the data features. According to Wang's research [11], the size of ECN plays a crucial role in affecting other ECN properties, so we show the distribution of the egonetwork sizes in the first place as in Fig. 1(a).

Group View. With the help of statistical view, it is easy to figure out great quantity users' contacts are in 200 or less. In order to explore users' distribute of the relationship about contacts number and call frequency, we develop the Group View with a chessboard layout, also we classify clusters under the density of users' distribute.

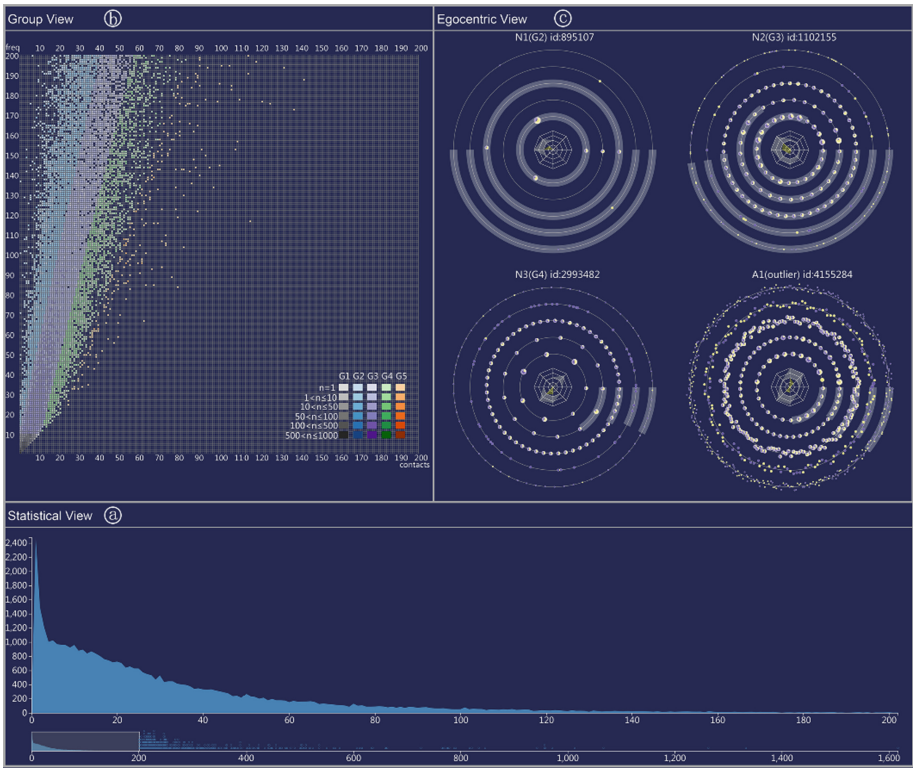

Fig. 1. (a) The visualization of users contacts and call frequency in 200 or less; (b) in this section, users communication structure in different clusters are represented as sky map to explore users features; (c) Monitoring users contacts in the entire telecommunications network.

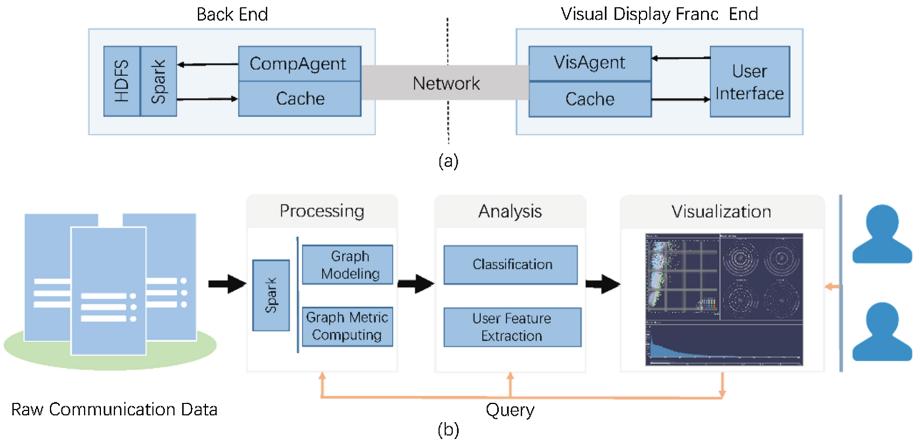

Fig. 2. The system architecture and data processing pipeline. (a) The system architecture; (b) The data processing pipeline.

With the classified clusters, we can figure out some specific user groups (one or several points in the view) we are interested in. In order to explore the distribute of users' coordinates, we map a series of sequence of colors to represent the number of users, refer to the bottom right corner in Fig. 1(b). The five clusters can be found in Fig. 1, as classified with the user density and the relationship of contacts and call frequency, and it can be seen that the "G1" has dark colors, also there are also many dark colors in the "G3", meanwhile, we interest in the "G2", some points have a few contacts with a quantity of frequency, with the "G4", covering the most points in right; we also interested in some points whose call frequency and contact ratio nearly equals 1. For further analysis, we display Egocentric View.

Egocentric View. With statistical view and group view, the analysts may interested in some specific users either for their representativeness or uniqueness. In fact, comprehending the social relationships and the communication strength of an ego can help to infer his/her personal social conditions. In order to fulfill these requirements and display the communication structure of users distinctly, this analytics system presents a microscopic user ego view with a sky map, which provides the following advantage: (1) Clearly display alien alters and local alters; (2) sky map layout have scalability to display the user with many contacts; (3) The layout display the character of the ego user clearly. In sky map, the alters which have a background is local user, the arc length of background represent the percent of the number of local alters, and the alter located in the near pathway means that the center user has more interconnection with the alter. and the inner ring displayed to compare the call frequency of center user with the alter, meanwhile, the outer ring compare the call duration of center user with the alter, and yellow displayed the called and purple represented the dialing. for the center, it takes a radar map to display the eight attributes of the central user which can be found in Fig. 3. Till now, the three views of the proposed visual analytics system have been fully presented in great details. The overall Statistical view provides the glimpse of all the users and their contacts distribute, and it helps the analysts promptly grasp the most users' property patterns. In order to get further information about the patterns located in the first view, the group view is proposed to compare egos within and without groups by applying glyph design. This view can help the analysts figure out interesting egos who need to be further investigated. Then the signature view presents the interactions between ego and alters as well as the interactions among the close friends. With all the above views, analysts are able to investigate the communication data from macroscopic, mesoscopic, and microscopic levels. This system can help the experts to know better about the social activity status of a specific user.

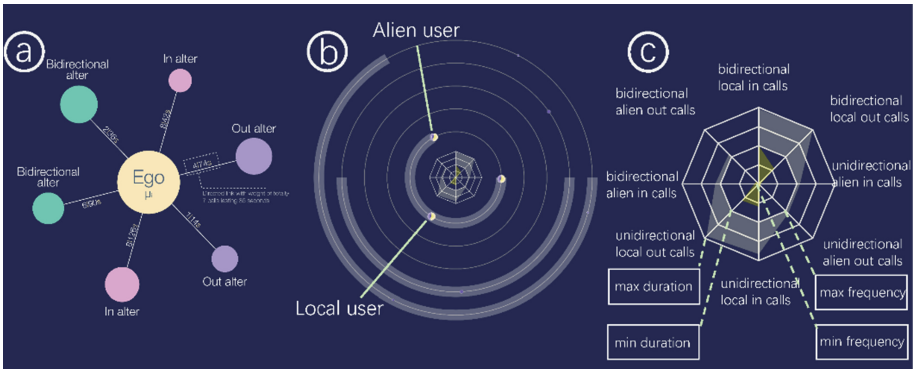

Fig. 3. Egocentric view: the source of data metric and the display of egocentric view. (a) the data metric of users; (b) This section put the design of the egocentric. (c) The detail information of ego center.

5 Case Study

In Fig. 1, the user "N1" from "G2" has few contacts but high calling frequency, and it interacts with its local alters at very high frequency. From the operator's perspective, it is a loyal user for its strong social relationships are all within this operator; There are lots of users in "G3", and "N2" is taken as an example, the user has intense communication with the bidirectional alters and lots of incoming calls from the other operators, thus it can be infered that it is a loyal user, meanwhile, it has the potential to attract the users from the other operators for it receives lots of calls from the other operators; For the users in "G4", like "N3", it has lots of contacts and has more interactions with local users, especially local bidirectional users (the center glyph), so this is an active and loyal user. All the above users have reasonable communication behaviors (both strong and weak social relationships are observed) and are normal users.

The last user "A1", in Fig. 1(c) has large number of contacts, however this user is outside of the scope of any existing group. In fact, this user calls lots of alien users, but doesn't have any strong relationships recorded, it looks more like an abnormal user, for example the telecom spammer. Figure 4 shows "A2" from "G1" and "A3" from "G5", according to definition, these are the users with very small and large ego networks, both of them are abnormal users. Among them, "A2" may have other mobile number in service at the same time, the egonetwork we obtained only contains part of its communication behaviors, and this is an alarm for the operator, and it calls for new business strategies to attract these users back. "A3" has 788 contacts (including many external contacts), the operator should try to maintain such heavy telecom user. From the above, we can see that this visual analytics system can help the operator to better understand the communication behaviors of both normal users and abnormal users, thus can help in making more personalized and profitable strategies.

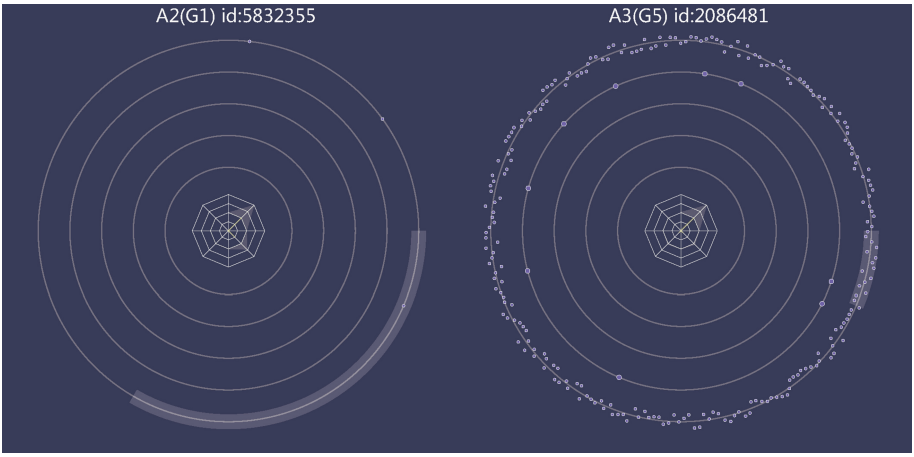

Fig. 4. In this section, we apply the visual analytics system in the task of abnormal user detection. The goal of this task is to find the abnormal users whose communication behaviors are very different and obtain the useful information to support the analysts of the mobile operator.

6 Discussion and Conclusion

The case study has demonstrated the efficacy of our visualization method in exploring large communication networks from egocentric perspective. The design of the statistical view can present the overall users' contacts distributions; the glyph based group view makes it easier to compare several users from the distribution of their contacts and call frequency at the same time; the stacked egocentric view can present the detail communication information of an ego. For example, this system can help to detect the service number and telecom-spammer.

Nevertheless, our method also suffers from several limitations. First of all, egocentric view can only show four egos at a time now. Secondly, the center of egocentric view high dimension information, and cannot display the ego's relationship network. The advantage is that such ratios can help the operator to estimate its market shares, and the disadvantage is the lost of the exact number of different kinds of alters.

As future works, the potential research direction is to present the ego-information of the ego and alters at the same time to improve mobility predictions.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant Nos. 61502083 and 61872066).

References

1. Manyika, J., et al.: Big data: the next frontier for innovation, competition, and productivity (2011). http://www.mckinsey.com/businessfunctions/business-technology/our-insights/big-data-the-nextfrontier-for-innovation. Accessed 10 Aug 2016

2. Barabási, A.L.: The origin of bursts and heavy tails in human dynamics. Nature **435**, 207–211 (2005)
3. Borgatti, S.P., Mehra, A.M., Brass, D.J., Labianca, J.: Network analysis in the social sciences. Science **323**, 892–895 (2009)
4. Roberts, S.G.B., Dunbar, R.I.M.: Communication in social networks: effects of kinship, network size, and emotional closeness. Pers. Relat. **18**, 439–452 (2011)
5. Onnela, J.P., et al.: Analysis of a large-scale weighted network of one-to-one human communication. New J. Phys. **9**, 179–206 (2007)
6. Eagle, N., Pentland, A.S., Lazer, D.: Inferring friendship network structure by using mobile phone data. Proc. Natl. Acad. Sci. U.S.A. **106**, 15274–15278 (2009)
7. Japkowicz, N., Stefanowski, J.: A machine learning perspective on big data analysis. In: Japkowicz, N., Stefanowski, J. (eds.) Big Data Analysis: New Algorithms for a New Society. SBD, vol. 16, pp. 1–31. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-26989-4_1
8. Hey, T.: The fourth paradigm – data-intensive scientific discovery. In: Kurbanoğlu, S., Al, U., Erdoğan, P.L., Tonta, Y., Uçak, N. (eds.) IMCW 2012. CCIS, vol. 317, p. 1. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33299-9_1
9. Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: definition, process, and challenges. In: Kerren, A., Stasko, J.T., Fekete, J.-D., North, C. (eds.) Information Visualization. LNCS, vol. 4950, pp. 154–175. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-70956-5_7
10. Mazza, R.: Introduction to Information Visualization. Springer, London (2009). https://doi.org/10.1007/978-1-84800-219-7
11. Wang, Q., Gao, J., Zhou, T., Hu, Z., Tian, H.: Critical size of ego communication networks. Europhys. Lett. **114**, 58004 (2016)
12. Onnela, J.P., et al.: Structure and tie strengths in mobile communication networks. Proc. Natl. Acad. Sci. U.S.A. **104**, 7332–7336 (2007)
13. Saramäki, J., Leicht, E.A., López, E., Roberts, S.G.B., Reed-Tsochas, F., Dunbar, R.I.M.: Persistence of social signatures in human communication. Proc. Natl. Acad. Sci. U.S.A. **111**, 942–947 (2014)
14. Shi, L., Wang, C., Wen, Z., Qu, H., Liao, Q.: 1.5D egocentric dynamic network visualization. IEEE Trans. Vis. Comput. Graphics **21**, 624–637 (2015)
15. Liu, Q., Hu, Y., Shi, L., Mu, X., Zhang, Y., Tang, J.: EgoNetCloud: event based egocentric dynamic network visualization. In: IEEE Conference on Visual Analytics Science and Technology, pp. 65–72 (2015)
16. Wu, Y., Pitipornvivat, N., Zhao, J., Yang, S., Huang, G., Qu, H.: EgoSlider: visual analysis of egocentric network evolution. IEEE Trans. Vis. Comput. Graphics **22**, 260–269 (2016)
17. Cao, N., Shi, C., Lin, S., Lu, J., Lin, Y., Lin, C.: TargetVue: visual analysis of anomalous user behaviors in online communication systems. IEEE Trans. Vis. Comput. Graph. **22**, 280–289 (2016)
18. Liu, D., Guo, F., Deng, B., Wu, Y., Qu, H.: EgoComp: a nodelink based technique for visual comparison of ego-network (2016). http://vacommunity.org/egas2015/papers/IEEEEGAS2015-DongyuLiu.pdf. Accessed 10 Aug 2016
19. Zhou, W.X., Sornette, D., Hill, R.A., Dunbar, R.I.M.: Discrete hierarchical organization of social group sizes. Proc. R. Soc. B **272**, 439–444 (2005)
20. Brzozowski, M.J., Romero, D.M.: Who should I follow? Recommending people in directed social networks. In: Fifth International AAAI Conference on Weblogs and Social Media, pp. 458–461 (2011)

21. Zhu, Y., Zhang, X., Sun, G., Tang, M., Zhou, T., Zhang, Z.: Influence of reciprocal links in social networks. PLoS One **9**, e103007 (2014)
22. Levandowsky, M., Winter, D.: Distance between sets. Nature **234**, 34–35 (1971)
23. Brown, J.J., Reingen, P.H.: Social ties and word-of-mouth referral behavior. J. Consum. Res. **14**, 350–362 (1987)
24. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I.: Spark: cluster computing with working sets. HotCloud **10**, 95 (2010)