ELSEVIER

CrossMark

Available online at www.sciencedirect.com

ScienceDirect

Physics Procedia 62 (2015) 36 – 41

Physics Procedia

3rd International Conference Frontiers in Diagnostic Technologies, ICFDT3 2013

Ranking Silent Nodes in Information Networks: a Quantitative Approach and Applications

Roberto Interdonato, Andrea Tagarelli

DIMES, University of Calabria, Rende (CS), Italy

Abstract

This paper overviews recent research findings concerning a new challenging problem in information networks, namely identifying and ranking silent nodes. We present three case studies which show how silent nodes' behavior maps to different situations in computer networks, online social networks, and online collaboration networks, and we discuss major benefits in identifying and ranking silent nodes in such networks. We also provide an overview of our proposed approach, which relies on a new eigenvector-centrality graph-based ranking method built on a silent-oriented network model.

Peer-review under responsibility of the ENEA Fusion Technical Unit

Keywords: silent-node behavior in information networks, ranking, eigenvector centrality, computer networks, social networks, collaboration networks

1. Introduction

Information networks are real-world complex systems that are characterized by highly interconnected informative units. Examples can naturally be found in a plethora of scenarios, including computer networks, biological and chemical systems, neural networks, online communities, and social networks (Boccaletti et al., 2006). In the last years, research in network analysis has spanned a variety of information retrieval and knowledge discovery tasks, such as community discovery, link prediction, keyword search, expert finding, and information diffusion related tasks. Particular attention has been devoted to the understanding of the behavior of nodes, and underlying relations, that are relevant to a specific task due to their "centrality" in the network; the latter is often expressed in terms of a notion of importance or influence, which in turn relies on the active roles the nodes play in the network, i.e., how and to what extent they produce or diffuse information.

However, an important aspect that has been neglected so far concerns the *passive* or *silent* role that is taken by nodes that do not readily inject information throughout the network. Such nodes, which typically represent the large majority in most information networks, remain quite unnoticed while benefiting from others' information or services without significantly giving back to the system. Identifying and ranking silent nodes within a network is a challenging task

E-mail addresses: rinterdonato@dimes.unical.it (Roberto Interdonato)., tagarelli@dimes.unical.it (Andrea Tagarelli).

Peer-review under responsibility of the ENEA Fusion Technical Unit
doi:10.1016/j.phpro.2015.02.008

as it in principle allows the system administrators to manage sub-networks particularly affected by this phenomenon, and ultimately to define personalized triggers in order to activate silent units.

In this paper, we discuss the novel problem of *ranking silent nodes* in information networks, which poses new challenges in network analysis and mining related areas. We begin with introducing three case studies which show how silent nodes' behavior maps to different situations in computer networks, online social networks and collaboration networks. We also discuss what are the major benefits in identifying and ranking silent nodes. We then present an overview of our proposed approach, which relies on a new PageRank-like graph-based ranking method built on a silent-oriented network model. We finally provide a summary of main experimental findings obtained by our research work so far.

2. Applications

We believe that the problem of identifying and ranking silent nodes finds application in a variety of information networks. Here we discuss three different instances of the silent-node ranking problem concerning the scenarios of computer networks, online social networks, and online scientific/professional collaboration networks. A common principle that characterizes silent nodes in any information network is that for those nodes there is often an unbalance between the information consumed with respect to the information produced.

2.1. Computer networks

Silent nodes in computer networks (e.g., P2P systems like BitTorrent, Gnutella) can be identified in the context of analysis of *leeching* behaviors.

Leech nodes act like parasites in absorbing others' information without giving anything back, or offering just the bare essential to access the network without being penalized or banned. This behavior may occur in several ways, depending on the type of computer network: accessing a wireless network without being allowed by the owner, selling content which was uploaded on the Internet as freely available by the rightful owners, receiving a reward in an online multi-player game (e.g., a MMORPG) without contributing to the team effort needed to achieve it. Nevertheless, the most common scenario associated to the leech figure is that of P2P networks. In a P2P network (Dhungel et al., 2008), a leech is a user downloading a huge amount of data without sharing anything, or disconnecting from the network as soon as s/he completes his downloads; however, on some P2P networks such as BitTorrent, leeching can also be associated to a legitimate practice of incomplete file-sharing. Note that a leeching behavior, despite not being generally illegal, is often considered malicious or harmful, as it usually violates the community's etiquette. Therefore, it's desirable to identify leeching nodes at various levels in order to design anti-leeching plans in the network.

2.2. Online social networks

The majority of members in online social networks can be considered silent users, as they exhibit a passive behavior and do not inject information in the shared online space. Such individuals are often called *lurkers*, since they remain quite unnoticed while benefiting from information or services offered by other members, without significantly giving back to the community (Nonnecke and Preece, 2000; Preece et al., 2004). Lurking is usually associated with definitions of nonparticipation, infrequent or occasional posting and, more generally, with observation, inactivity/passivity, and bystander behavior (Lave and Wenger, 1991; Kahnwald and Kohler, 2006; Halfaker et al., 2013). Lurking can be expected or even encouraged because it allows users to learn or improve their understanding of the etiquette of an online community before they can decide to provide a valuable contribution over time (Edelmann, 2013); ultimately, they can be attracted by online advertising strategies tailored to specific lurkers' behavioral profiles with the ultimate objective of *de-lurking* them.

Despite the fact that lurkers represent the large majority of members in a social network — it has been estimated that at any point in time approximately 90% of community members may be lurkers (Nonnecke and Preece, 2000) — little research has been done that considers lurking as a valid and worthy-of-investigation form of online behavior. Identifying lurkers represents a great potential to the understanding of social dynamics in the network. In particular, ranking lurkers can be useful to manage priority in de-lurking applications, to identify the sub-communities particularly affected by lurkers, and to define personalized triggers of active participation (Tagarelli and Interdonato, 2013b).

2.3. Collaboration networks

Collaboration networks (CNs) are prototypes of information networks constituted by (possibly) heterogeneous entities (e.g., organizations, people, projects, scientific publications, encyclopedic entries) which interact with each other in order to achieve common or compatible goals. A typical example of CNs is represented by research collaboration networks. Here, silent nodes can be regarded as nodes with "non-expert" roles, that is, *apprentices* or *advisees*. Like for online social networks, such nodes constitute a significant part of members in the CN, since an apprenticeship status clearly holds for the initial stage of the researcher lifetime, and also with respect to any topic that at a particular time does not represent the researcher's interests.

A particularly challenging type of relationship to discover from the apprentice perspective concerns *vicarious learning*. While in social learning theory this definition assumes a positive meaning (e.g., people can learn through being given access to the learning experiences of others (Bandura, 1986)), in a publication context it can still be identified and measured in collaborations in which one might marginally contribute to the research activity (Tagarelli and Interdonato, 2013a).

Another prominent example of CNs is that of Wiki-edit networks, which represent communities behind collaboratively edited encyclopedias (e.g., Wikipedia, Wiktionary). A major focus here is on the relations between editors of wikis. In this context, silent nodes are those editors who offer occasional or scarcely relevant contributions to the increase and improvement of the encyclopedic entries. Therefore, identifying and ranking them can aid to foster gamification-like awards for users who are willing to increase their productivity.

3. A silent-oriented network model

Any information network can be conveniently represented as a graph, whose nodes are entities (e.g., users, computing hosts, items, etc.) and links model relations between entities (i.e., communication, shared information, other forms of human interaction) (Wasserman and Faust, 1994). Relations can be symmetric or asymmetric; in the latter case, node interactions are typically modeled in such a way that the role of center of mass is played by nodes that are the most important in terms of information propagation across the network. According to classic eigenvector-centrality graph-based ranking methods, such as PageRank (Brin and Page, 1998) and alpha-centrality (Bonacich and Lloyd, 2001), the more (or more relevant) incoming links a node has the more important it is. This concept of importance or centrality is often associated to a notion of influence or authoritativeness of nodes in the network, hence we refer here to such a type of graph as influence-oriented graph.

However, if the focus is on silent nodes' behavior, then the way a node should be regarded as central or important needs to be revised. An intuitive way would be changing the edge orientation from the node who produces information to the node that receives it. In our previous works (Tagarelli and Interdonato, 2013b,a), we have indeed shown that by reversing the network's influence-oriented topology, the resulting graph model fits well the principle that the greater the amount of information a node receives, and contemporarily the smaller the amount of information produced by that node, then the higher the probability of being regarded as a silent node.

Unfortunately, straightforward application of PageRank or related methods on the reversed, silent-oriented network does not imply an inversion too of top-ranked positions in favor of silent nodes (Tagarelli and Interdonato, 2013b,a). By contrast, topological indicators of silent nodes' behavior are needed to define an effective ranking method.

4. Ranking of silent nodes in an information network

In order to score a silent node in the network, our first principled step is to model the mutual contribution from incoming and outgoing links through the node's in/out-degree ratio, so that the higher this ratio, the higher the likelihood for the node to be silent (vice versa, a relatively high out/in-degree ratio would hint that the node is influential). However, the in/out-degree ratio is not a sufficient criterion to identify and rank silent nodes: the authoritativeness of neighbors is not taken into account and, moreover, there would not be enough diversity in the ranking scores (e.g., many nodes would obtain the same or very close ranking scores).

Our key idea to design an effective ranking method for silent nodes lies in better leveraging a node's incoming and outgoing connections. Based on this intuition, the strength of the silent status of a node should also be determined proportionally to the strength of non-silent behavior shown by its in-neighbors, and to the strength of silent behavior shown by its out-neighbors. The following is a definition of *topology-driven silent-node behavior*, which is adapted from (Tagarelli and Interdonato, 2013b).

Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ denote the directed graph representing an information network, with set of nodes (members) $\mathcal{V}$ and set of edges $\mathcal{E}$, whereby the semantics of any edge (u, v) is that v is receiving or consuming information from u. A node v with infinite in/out-degree ratio (i.e., a sink node) is trivially regarded as a silent node. A node v with in/out-degree ratio above 1 shows a silent status, whose strength is determined proportionally to (*i*) the in/out-degree ratio, (*ii*) the strength of non-silent behavior shown by in-neighbors of v, and (*iii*) the strength of silent behavior shown by out-neighbors of v.

According to the above definition, we define a *silent-node ranking* measure as follows:

$$r_i = \left(\frac{1}{outdeg(i)} \sum_{j \in In(i)} \frac{outdeg(j)}{indeg(j)} r_j \right) \left(1 + \left(\frac{indeg(i)}{\sum_{j \in Out(i)} indeg(j)} \sum_{j \in Out(i)} \frac{indeg(j)}{outdeg(j)} r_j \right) \right) \tag{1}$$

where $indeg(\cdot)$ (resp. $outdeg(\cdot)$) is the add-one smoothed in-degree (resp. out-degree) function, $In(i)$ (resp. $Out(i)$) is the set of in-neighbors (resp. out-neighbors) of node i, and r_j is the ranking score of node j. It can be noted the above formula has two main terms. Intuitively, due to the first term, the score of node i increases with the number of its in-neighbors and with their likelihood of being non-silent nodes, which is expressed by a relatively high out/in-degree. The first term also includes a factor that is inversely proportional to the i's out-degree, in order to take into account both the contribution of a node's in-neighbors and its own in/out-degree property. The second term in the formula lets the score of a node increase with the tendency of its out-neighbors of being silent, and includes a correction factor proportional to the in-degree of the target node, in order to score i higher if it receives more than what its out-neighbors receive.

A complete specification of our silent-node ranking measure can be provided by resorting to an eigenvector-centrality scheme, like that provided by PageRank and alpha-centrality; here, for the sake of brevity, we refer to our PageRank-based formulation, while the interested reader can find details in (Tagarelli and Interdonato, 2013b). Recall that the PageRank vector is the unique solution of the iterative equation $\mathbf{r} = \alpha \mathbf{S} \mathbf{r} + (1 - \alpha)\mathbf{p}$, where $\mathbf{S}$ denotes the column-stochastic transition probability matrix, $\mathbf{p}$ is a probability distribution vector, and α is a real-valued coefficient in [0,1] which acts as damping factor. Our proposed PageRank-based silent-node ranking score, for any node i, is defined as follows:

$$r_i = \alpha \left[\left(\frac{1}{outdeg(i)} \sum_{j \in In(i)} w(j,i) \frac{outdeg(j)}{indeg(j)} r_j \right) \left(1 + \left(\frac{indeg(i)}{\sum_{j \in Out(i)} indeg(j)} \sum_{j \in Out(i)} w(i,j) \frac{indeg(j)}{outdeg(j)} r_j \right) \right) \right] + \frac{1 - \alpha}{|\mathcal{V}|} \tag{2}$$

where $w(\cdot, \cdot)$ is the edge weighting function.

5. Experimental Findings

To asses the significance of our approach to the ranking of silent nodes, we conducted a number of quantitative and qualitative experimental analyses in the domain of online social networks (Tagarelli and Interdonato, 2013b) and collaboration networks (Tagarelli and Interdonato, 2013a). We will now summarize our main experimental findings.

5.1. Lurker ranking

For the evaluation on the online social network domain, we used information from *Twitter* and *FriendFeed*. Our analysis on Twitter was conducted using a subset of the dataset studied in (Kwak et al., 2010), containing about 16 million users and 132 million links. As regards FriendFeed, we used the latest version of the dataset studied in (Celli et al., 2010), which consists of about 500 thousand users and 19 million links. For the quantitative analysis we focused on the comparison of various formulations of our Lurker Ranking method with respect to several competing methods (PageRank, alpha-centrality, and Fair Bets (Budalakoti and Bekkerman, 2012)) as well as with respect to a *data-driven ranking* (i.e., a simulated ground-truth based on social network-specific measures of influence). For this evaluation we resorted to well-known assessment criteria, namely *Fagin's intersection metric* (Fagin et al., 2003) and *Bpref* (Buckley and Voorhees, 2004). The formulation described in Section 4 (together with the in-neighbors based one, which takes into account only the first term in equation 1) obtained the best performances on both datasets, showing high correlation with the data-driven ranking and poor correlation with the competing methods. As concerns the qualitative analysis, we compared the top-20 lurkers as ranked by our method and by the competing methods, analyzing the social network profiles of the corresponding users. Results revealed how the top-ranked list of the competing methods contained several influential users (active people with lots of followers and retweets), while the top-ranked list of our method only contained users that were recognized as lurkers (silent nodes), thus demonstrating the effectiveness of our solution.

5.2. Vicarious learner ranking

Our experimentation in the collaboration network domain was focused on a task of identification and ranking of vicarious learners. We used an XML dump of the DBLP computer science bibliography.[1] We extracted information about the number of joint publications for each pair of co-authors, and the total number of publications for every author, on a yearly basis. We evaluated our method on both the complete DBLP dataset (more than 1 million nodes and about 5 million links) and on three subsets corresponding to the last three terms of approximately three years (in order to study the temporal evolution of the identified vicarious learners). We used PageRank as competing method, and evaluated the performance against two *data-driven rankings*: one using information from DBLP itself (e.g., number of coauthors, number of single-authored publications) and an alternative one consisting of a ranking based on the *activity score* extracted from the ArnetMiner website.[2] We used the *Kendall tau rank correlation coefficient* (Abdi, 2007) to measure the correlation of the ranking results with respect to the DBLP-based and the ArnetMiner-based reference rankings, and we found out that our method always outperformed PageRank in terms of correlation with both reference rankings. Effectiveness of our method was confirmed by a qualitative analysis on the top-100 ranked list by the two algorithms: our method was able to assign highest scores to authors who can be defined vicarious learners with a certain objectivity, while in the top-ranked list by PageRank we also found authors who are more likely to be defined as team leaders or at least active contributors. Our method showed to be more effective than PageRank also in capturing the temporal evolution of vicarious learners, as demonstrated by the results of the experiments on the three-year subsets of DBLP.

6. Conclusions and Future Work

Identifying and ranking silent nodes in information networks is a challenging problem in network analysis and mining. Our proposed approach exploits the network topology information to define a new eigenvector-centrality graph-based method that is well-suited to detect and score silent nodes.

The inherent complexity and variety of contexts in which silent nodes' behaviors can be recognized, however advises that other information than the network topology might be used for a better understanding and analysis. Specifically, including temporal aspects (e.g., online access frequency) is important to analyze the evolution of network members' roles over time, and then to study the causes of a stronger or weaker tendency toward silent behavior on different times. Another way to improve silent behavior analysis should be based on exploiting contextual information, such as the type of resources exchanged, the topics discussed, or the work tasks shared among the network units (e.g., community members). This would suggest an analysis based on the context-biased activity of silent nodes, e.g., studying how the same node can be involved in different aspects underlying the relationships in the network.

[1] http://dblp.uni-trier.de/db/
[2] http://arnetminer.org/AcademicStatistics

References

Abdi, H., 2007. The Kendall Rank Correlation Coefficient, in: Encyclopedia of Measurement and Statistics.

Bandura, A., 1986. Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice Hall.

Boccaletti, S., Latora, V., Moreno, Y., M. Chavez, D.U.H., 2006. Complex networks: Structure and dynamics. Physics Reports 424, 175–308.

Bonacich, P., Lloyd, P., 2001. Eigenvector-like measures of centrality for asymmetric relations. Social Networks 23, 191–201.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30, 107–117.

Buckley, C., Voorhees, E.M., 2004. Retrieval evaluation with incomplete information, in: Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), pp. 25–32.

Budalakoti, S., Bekkerman, R., 2012. Bimodal invitation-navigation fair bets model for authority identification in a social network, in: Proc. ACM Conf. on World Wide Web (WWW), pp. 709–718.

Celli, F., Lascio, F.M.L.D., Magnani, M., Pacelli, B., Rossi, L., 2010. Social Network Data and Practices: The Case of Friendfeed, in: Proc. Int. Conf. on Social Computing, Behavioral Modeling, and Prediction (SBP), pp. 346–353.

Dhungel, P., Wu, D., Schonhorst, B., Ross, K.W., 2008. A measurement study of attacks on BitTorrent leechers, in: Proc. Conf. on Peer-to-peer systems (IPTPS), p. 7.

Edelmann, N., 2013. Reviewing the definitions of "lurkers" and some implications for online research. Cyberpsychology, Behavior, and Social Networking 16, 645–649.

Fagin, R., Kumar, R., Sivakumar, D., 2003. Comparing Top k Lists. SIAM Journal on Discrete Mathematics 17, 134–160.

Halfaker, A., Keyes, O., Taraborelli, D., 2013. Making peripheral participation legitimate: reader engagement experiments in Wikipedia, in: Proc. ACM Conf. on Computer Supported Cooperative Work (CSCW), pp. 849–860.

Kahnwald, N., Kohler, T., 2006. Microlearning in virtual communities of practice? an explorative analysis of changing information behaviour, in: Proc. Microlearning Conf., pp. 157–172.

Kwak, H., Lee, C., Park, H., Moon, S.B., 2010. What is Twitter, a social network or a news media?, in: Proc. ACM Conf. on World Wide Web (WWW), pp. 591–600.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press.

Nonnecke, B., Preece, J.J., 2000. Lurker demographics: counting the silent, in: Proc. ACM Conf. on Human Factors in Computing Systems (CHI), pp. 73–80.

Preece, J.J., Nonnecke, B., Andrews, D., 2004. The top five reasons for lurking: improving community experiences for everyone. Computers in Human Behavior 20, 201–223.

Tagarelli, A., Interdonato, R., 2013a. Ranking Vicarious Learners in Research Collaboration Networks, in: Urs, S., Na, J.C., Buchanan, G. (Eds.), Digital Libraries: Social Media and Community Networks. Springer International Publishing. volume 8279 of *LNCS*, pp. 93–102.

Tagarelli, A., Interdonato, R., 2013b. "Who's out there?": identifying and ranking lurkers in social networks, in: Proc. IEEE/ACM Conf. Advances in Social Networks Analysis and Mining (ASONAM), pp. 215–222.

Wasserman, S., Faust, K., 1994. Social Networks Analysis: Methods and Applications. Cambridge University Press.