Investigation of Localised Centrality Metrics for Collaborative Networks: What can they reveal?

Elizabeth M. Daly

IBM, Dublin Software Lab
IBM Software Group
Elizabeth_Daly@ie.ibm.com

Abstract. Collaborative web 2.0 applications, such as blogs, collaborative bookmarking, file sharing etc., have increased significantly in popularity. In these user-centric applications users are not only consumers, but also contributors. By contributing content to the system, users become part of the network and relationships between users and content can be derived. Social network metrics can be used to identify key users, however, evaluating network metrics for a large scale network can be expensive. For this reason this paper explores the utility of localised network metrics. Experimental results and analysis are presented on a large collaborative IBM bookmarking network called Dogear to investigate the ability to identify central users.

1 Introduction

Web 2.0 is a new medium where users are not just consumers, but are also contributors. By contributing content to applications, such as CiteULike, Delicious or Digg, users become part of the network and relationships between users can be derived forming a social network. The corporate world has started to deploy these types of colloborative applications in order to promote connectivity and, by extension, encourage innovation and productivity. In an organisation, an employee's connections can represent resources they have access to. Evaluating the network of user relationships can aid in determining what role the user plays in knowldge networks. By doing so organisations could better learn how to identify succession candidates, potential bottlenecks and over dependencies, improve cross boundary communication and collaboration. Importantly, the structure of these informal knowledge networks may bare little resemblance to an organisational hierarchy [12].

Cross and Parker define three key roles [4]; *Central connectors* who have a large number of contacts and can represent people who may play an important connecting role with access to a large number of resources. If these users are highly connected, they may also represent bottlenecks where they are overloaded, additionaly, the important role these users play may also have gone unnoticed. The network may be heavily dependent on these users for connectivity and what happens if these people retire or leave? *Boundary spanners* who provide critial links between groups of people that may be seperated through functional roles, geography etc. The diversity of a user's type of contacts can also be seen as a user having access to a larger set of non-redundant social

resources. *Information brokers* as highlighted by Granovetter who argued the utility of using weak ties for information flow in social networks [9]. He emphasised that weak ties lead to information dissemination between groups. He introduced the concept of 'bridges', observing that 'those who are weakly tied are more likely to move in circles different from our own and will thus have access to information different from that which we receive' [9].

Various centrality metrics exist to aid in identifying these different types of users, however collaborative networks can consist of thousands if not millions of users, as networks become larger the computation of global network metrics becomes increasingly expensive. In such cases, the option of exploring local networks can be beneficial [5]. For this reason, this paper explores the utility of localized network metrics in a collaborative network. Experimental results are presented on a collaborative IBM bookmarking network called Dogear investigating what these metrics reveal.

2 Related Work

Due to the interactive nature of web 2.0 applications, they provide a rich source for analysing user behaviour and interrelationships. The social networks that arise have been the subject of much recent research on mining social relationships. Relationships between users can be explicit such as friending in Facebook, or implicit through users having similar interests, content and tagging behaviour.

Lewis et al. use facebook data limited to a group university students in order to gain insights into different user online behaviour based on features such as gender and ethnicity [13]. Additionally, they also examine the different types of online ties such as friendship ties, shared photo ties and college housing ties. Kolari et al. examined the structure and utility of an IBM internal blog network [11]. The authors investigate network structure, cross domain interaction and impact metrics taking into account corporate hierarchy. Bakshy et al. use second life data to model the transfer of user generated content and that adoption is related to whether the content comes from a friend, or whether the content is adopted by those in a user's social circle [1]. Leydesdorff applies betweenness centrality as an indicator of the 'Interdisciplinarity' for scientific journals in a scientific citation network [14]. Golder and Huberman provide analysis of the tagging behavior and tag usage in online communities [8]. The authors provide an overview about the structure of collaborative tagging systems. Based on a small subset of the Delicious corpus, they investigate what motivates tagging and how tagging habits change over time. Chi and Mytkowicz investigate the dynamics of tags related to documents and shows that the information gained from a tag becomes less useful as the proliferation of use increases [3]. Bhattacharyya et al. use similarity of keywords in profiles to build a social graph in order to create a synthetic social graphs [2].

Anaysis of the social structure of collaborative applications can provide key insights into user behaviour, the influence users have on each other and the role of users and their relationships in these networks.

3 Localised Centrality Metrics

Node centrality is used to identify important nodes in a network. Centrality in graph theory and network analysis is a quantification of the relative importance of a vertex within the graph (e.g., how important a person is within a social network). A central node, typically, has a stronger capability of connecting other network members. There are several ways to measure centrality. The three most widely used centrality measures are Freeman's degree, closeness, and betweenness measures [6,7].

Freeman's centrality metrics are based on analysis of a complete and bounded network which is referred to as a sociocentric network. These metrics require evaluation of the entire network resulting in computationally expensive operations, which has motivated the introduction of 'ego networks'. An ego network can be defined as a network consisting of a single actor (ego) together with the actors they are connected to (alters) and all the links among those alters. Consequently, ego network analysis can be performed locally by individual nodes without complete knowledge of the entire network. Marsden introduces centrality measures calculated using ego networks and compares these to Freeman's centrality measures of a sociocentric network [15].

'Degree' centrality is measured as the number of direct ties that involve a given node [7]. A node with high degree centrality maintains contacts with numerous other network nodes. Such nodes can be seen as popular nodes with large numbers of links to others. As such, a central node occupies a structural position (network location) that may act as a conduit for information exchange. In contrast, peripheral nodes maintain few or no relations and thus are located at the margins of the network. Degree centrality for a given node p_i is calculated as:

$$C_D(p_i) = \sum_{k=1}^{N} a(p_i, p_k) \tag{1}$$

where $a(p_i, p_k) = 1$ if a direct link exists between p_i and p_k and $i \neq k$. Degree centrality can easily be measured for an ego network where it is a simple count of the number of contacts.

'Closeness' centrality measures the reciprocal of the mean geodesic distance $d(p_i, p_k)$, which is the shortest path between a node p_i and all other reachable nodes [7]. Closeness centrality can be regarded as a measure of how long it will take information to spread from a given node to other nodes in the network [18]. Closeness centrality is uninformative in an ego network, since by definition an ego network only considers nodes to which the ego node is directly related to and then by definition the distance from the ego node to all other nodes considered in the ego network is 1 and so is not included in this paper.

'Betweenness' centrality measures the extent to which a node lies on the shortest paths linking other nodes [6,7]. Betweenness centrality can be regarded as a measure of the extent to which a node has control over information flowing between others [18]. A node with a high betweenness centrality has a capacity to facilitate interactions between the nodes that it links. In our case it can be regarded as how well a node can facilitate communication to other nodes in the network. Betweenness centrality is calculated as:

$$C_B(p_i) = \sum_{j=1}^{N} \sum_{k=1}^{j-1} \frac{g_{jk}(p_i)}{g_{jk}} \tag{2}$$

where g_{jk} is the total number of geodesic paths linking p_j and p_k, and $g_{jk}(p_i)$ is the number of those geodesic paths that include p_i.

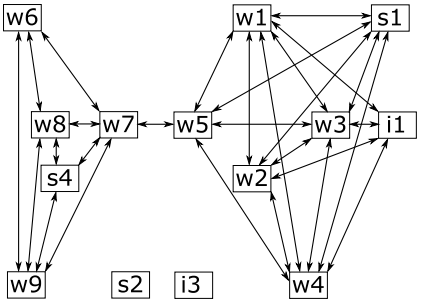
'Bridging' centrality has been more recently introduced as a measure of how much a node is located between highly connected regions [10]. The authors define a 'bridging coefficient' to capture the bridging behaviour of the nodes' neighbourhood. The aim is to capture nodes where a high amount of information may flow through, due to the density of the information available versus the number of contacts. The 'bridging coefficient' $BC(p_i)$ is given as:

$$BC(p_i) = \frac{\frac{1}{C_D(p_i)}}{\sum_{k=1}^{N} \frac{1}{C_D(p_k)}} \tag{3}$$

where N is the neighbourhood contacts of node p_i. Bridging centrality is then calculated as the product of the 'betweenness' centrality and the 'bridging coefficient'.

$$C_{Br}(p_i) = C_B(p_i) \times BC(p_i) \tag{4}$$

Betweenness centrality in ego networks has shown to be quite a good measure when compared to that of the sociocentric measure. Marsden calculates the egocentric and the sociocentric betweenness centrality measure for the network shown in figure 1 [15]. A more recent analysis extended this work by measuring an egocentric bridging centrality, which is also included in figure 1 .



Node	Sociocentric betweenness	Egocentric betweenness	Sociocentric bridging	Egocentric bridging
w1	3.75	0.83	0.528	0.117
w2	0.25	0.25	0.052	0.052
w3	3.75	0.83	0.528	0.117
w4	3.75	0.83	0.528	0.117
w5	30	4	6.667	0.889
w6	0	0	0	0
w7	28.33	4.33	5.074	0.775
w8	0.33	0.33	0.073	0.073
w9	0.33	0.33	0.073	0.073
s1	1.5	0.25	0.333	0.055
s2	0	0	0	0
s4	0	0	0	0
i1	0	0	0	0
i3	0	0	0	0

Fig. 1. Bank Wiring Room network sociocentric and egocentric betweenness

The betweenness centrality $C_B(p_i)$ based on the egocentric measures does not correspond perfectly to the sociocentric measures. However, it can be seen that the ranking of nodes based on the two measures of betweenness are identical in this network. In effect, the betweenness value captures 'how much a node connects nodes that are themselves

not directly connected'. Marsden compared sociocentric and egocentric betweenness for 15 other sample networks and found that the two values correlate well in all scenarios [15]. Similary, analysis by Nanda and Kotz found correlation between sociocentric and egocentric bridging centrality [17].

4 Experimental Results

In order to investigate the potential of localised centrality metrics a real world data set is analysed in order to determine what types of users these metrics uncover. IBM's collaborative bookmarking solution Dogear [16] is popularly used by IBM employees and the data set contains an extensive network of users and contributed URLs, shown in table 1. Egocentric betweenness is calculated for a symmetric relationship matrix using the computationally efficient method presented by Everett and Borgatti [5].

Table 1. Dogear dataset

Number of Users	Number of Bookmarks	Number of URLs
10259	505472	317362

4.1 URL relationship network

To gain an initial understanding of the structure of the collaborative bookmarking network, we first derive a simple URL relationship network. Users are linked through overlapping URLs contained in their document collection, if two users share a URL in their collection, then an edge exists between those users. Figure 2 a) shows a log-log plot of the distribution of the number of URLs users have in their collection. As expected the distribution follows the power law with a large number of users who only have a small number of bookmarks and a small group of power users. Figure 2 b) shows a log-log plot of the degree centrality distribution which similarly follows the power law.

Next we examine the relationship between the number of bookmark URLs a user has in their collection and the degree centrality. This demonstrates that the more a user participates in the network, the more links between other nodes are created.

4.2 Subscription network

The URL relationship network could be useful to infer implicit relationships, however, the Dogear dataset provides a subscription network where users subscribe to specific users' bookmark collections. This can be seen as an explicit identification of interest and can be used to validate whether the localised centrality metrics can identify important users in the network. A total of 1727 users participate in the subscription network either as subscribers or users that are subscribed to. In addition to the subscription information, the user's email was available to determine the domain each user belongs to. Therefore we can see when a user is highly connected within their domain, or whether their connections span geographic boundaries.

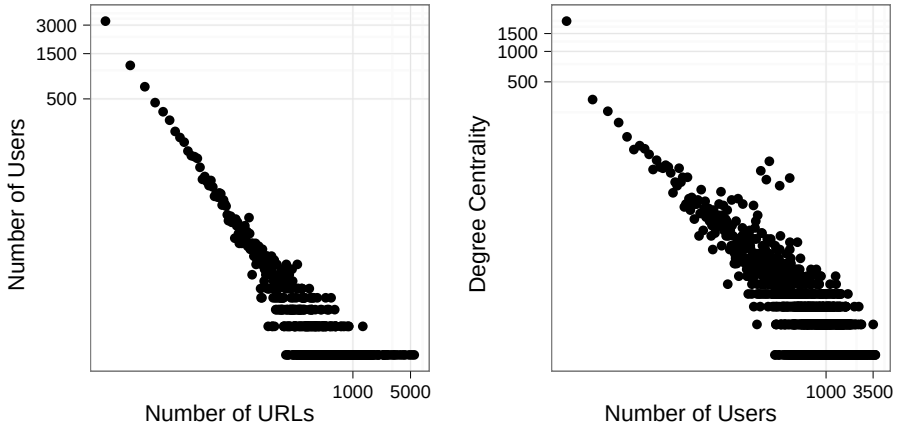


Fig. 2. URL relationship network - a) distribution of URLs per user b) degree centrality

Betweenness Centrality Figure 3 a) shows the degree centrality versus the number of domains the user spans. The node sizes represent the egocentric betweenness centrality. From the plot it seems clear that a high egocentric betweenness (shown as size of dot) results from a high number of contacts spanning a large number of domains. However, upon inspection, two nodes, node A and node B have a similar number of contacts, a similar betweenness value, and yet differ greatly on the number of domains they span. This can be explained when examining the variance of the social properties of each node shown in table 2. Out of the 46 contacts in node A's network, nearly 70% are bi-directional reciprocated links. In the subscription network the average percentage of a user's contacts that are reciprocated is less than 7%. When mapping the user back to an email domain, it turns out node A is in the Japanese mail domain where arguably the language barrier causes a higher interconnectivity. Node B has a similar number of contacts, however, it is connected to a large number of domains. In contrast to node A, this node has relatively few reciprocated contacts. From this, we can infer that one user is important for connecting users within their domain, and the other node is structurally important for connecting users across domains. Upon investigation, the user represented by Node B has since left the company and was involved in strategy and enabling business partnerships. Node C has a high value across all measures and spans 18 domains where the global average is less than two. Additionally, 97% of the user's contacts are out-links where users have subscribed to their bookmark feed, meaning this node may potentially influence many users. When mapping this node back to a specific user, this user is a well respected research fellow within IBM. Node D has also since left the company and played a key role in educating customers and employees about collaboration software. Node E has a large number of incoming and outgoing links and connects users across many domains. This user is well known in IBM as a prolific blogger who specialises in knowledge management. The important aspect to note, is that the betweenness metric captures users that have a useful social network either through the

domains they span, the strength of their ties, or the extent to which they are followed by others spanning domains. Additionally, two out of the five users identified of interest have since left the company. This highlights the importance of uncovering the roles different users play in the social graph in order to identify potential gaps their departure could create.

Table 2. Selected metrics related to betweenness of subscription network

Node id	degree centrality	domains	ego betweenness	out-degree	in-degree	reciprocated
A	46	1	749	41	39	32
B	49	11	771	42	10	3
C	192	18	14833	188	17	10
D	160	23	10985	148	26	9
E	134	25	8007	77	82	25
average	4.1	1.6	38.5	2.28	2.26	0.4

Bridging Centrality Figure 3 b) shows the same graph as figure 3 a), however, the sizing of the nodes are determined by node bridging centrality. As can be seen nodes with a high bridging centrality, have an inverse relationship to the number of contacts and number of domains they span. Nodes with a high bridging centrality span very few domains and have relatively few contacts in their social graph.

Bridging centrality highlights 4 nodes, node F and three other nodes with the same bridging centrality value 26.66, only node F is visible due to the overlap. Node F is shown in figure 4 and has only two contacts, node D and node G, the degree centrality of the neighbouring nodes are shown in table 3 which have a degree centrality of 160

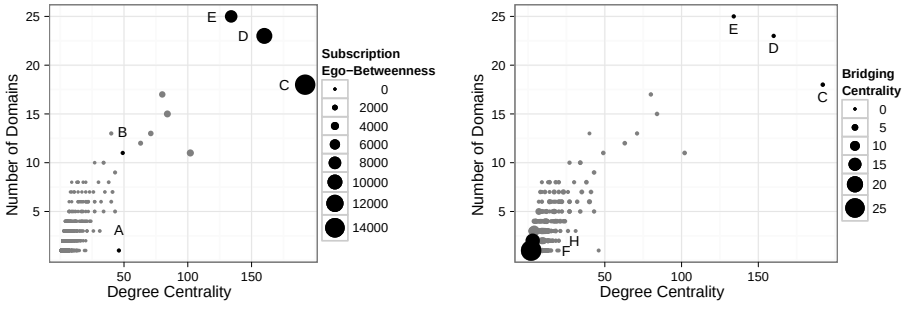


Fig. 3. Subscription network - Number of domains in user's network vs number of contacts (degree centrality) a) Betweenness centrality node sizing b) Bridging centrality node sizing

and 80 respectively. As can be seen these two nodes represent nodes with a high betweenness value and these two nodes are ranked 2nd and 6th in the network in terms of betweennness centrality. Interestingly, the other three nodes with the same betweenness centrality all link to the same two nodes. Upon inspection of the users, all four users are involved in technical sales and consulting roles. Node H has 3 contacts where 1 relationship is reciprocated. Table 3 shows that the neighbouring nodes have a degree centrality of 160, 134 and 40 respectively. Interestingly, Node H is involved in learning develoopement and identifies themselves as an early adopter of technology in their company profile. Bridging centrality has identified users with a small number of contacts that are connected to nodes with a high betweenness centrality. These nodes potentially have access to a dense amount of information using a relatively small number of contacts.

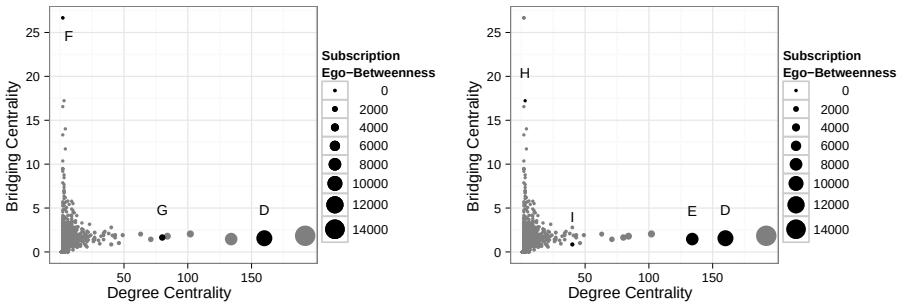


Fig. 4. Subscription network - Bridging centrality vs Degree centrality a) Node F's contacts b) Node H's contacts

Table 3. Selected metrics related to betweenness

Node id	degree centrality	domains	bridging centrality	in-degree	contacts degree centrality
F	2	1	26.66	2	160,80
H	3	2	17.22	3	160,134,40

4.3 Discussion

The advent of web 2.0 collaborative applications encourages users to interact and participate in online networks. Users tend to have different roles in the network, some are innovators and leaders, some are early adopters that closely follow new trends, others are highly focused and are only interested in a small subset of users and content. Many applications benefit from being able to identify a user's roles in the network.

In corporate environments that can span many geographic boundaries, enabling collaboration is important for knowledge transfer as contacts to individuals can represent access to knowledge. Betweenness centrality has been shown to be able to identify users that are key to the network graph in terms of information flow, linking many different users. Bridging centrality focuses on nodes that are not necessarily highly central for information flow, but are connected to highly between nodes and have a more focused, concentrated number of contacts.

5 Conclusion and Future Work

This paper has investigated the types of nodes that are identified in a collaborative bookmarking network using localised centrality metrics. Using real world data from the Dogear collaborative bookmarking application, we constructed two relationship graphs. A URL relationship graph linking users with overlapping bookmarks which we consider an implicit relationship, and a subscription relationship graph which we consider an explicit relationship. The analysis showed that betweenness centrality identifies users that have a rich social network either through the domains they span, the reciprocated nature of their ties, or the extent to which they are followed by others spanning domains. The subscription network is a small subset of the application users, as a result, other relationships must be used to provide a more complete insight into the network. The URL relationship network is one example, however the graph presented in this paper is symmetric, where there has been no identification of link directionality. One solution may be to assign direction to the links based on the date when the user adds the bookmark, in an attempt to identify the contributor and the consumer in the relationship. Another potential relationship is one based on overlapping use of tags in order to take into account the different types of relationships.

References

1. Eytan Bakshy, Brian Karrer, and Lada Adamic. Social influence and the diffusion of user-created content. In *10th ACM Conference On Electronic Commerce*. ACM, July 2009.
2. Prantik Bhattacharyya, Ankush Garg, and Felix S. Wu. Social network model based on keyword categorization. In *In ASONAM '09: Proceedings of the International Conference on Advances in Social Network Analysis and Mining,*, July 2009.
3. Ed H. Chi and Todd Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 81–88, New York, NY, USA, 2008. ACM.
4. Robert L. Cross, Andrew Parker, and Rob Cross. *The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations*. Harvard Business School Press, June 2004.
5. M. Everett and S. P. Borgatti. Ego network betweenness. *Social networks (Soc. networks)*, 27(1):31–38, 2005.
6. L. C. Freeman. A set of measures of centrality based on betweeness. *Sociometry*, pages 35–41, 1977.
7. L. C. Freeman. Centrality in social networks conceptual clarification. *Social networks (Soc. networks)*, pages 215–239, 1979.

8. Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, April 2006.
9. Mark S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, May 1973.
10. W. Hwang, Cho, and M. Ramanathan. Bridging centrality: Identifying bridging nodes in scale-free networks. Technical report, March 2006.
11. Pranam Kolari, Tim Finin, Kelly Lyons, Yelena Yesha, Yaacov Yesha, Stephen Perelgut, and Jen Hawkins. On the structure, properties and utility of internal corporate blogs. In *In International Conference on Weblogs and Social*, 2007.
12. Valdis Krebs. Managing the 21st century organization. volume xi. *International Association for Human Resource Information Management Journal*, XI, 2007.
13. Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using facebook.com. *Social Networks*, In Press, Accepted Manuscript, 2008.
14. Loet Leydesdorff. "betweeness centrality" as an indicator of the "interdisciplinarity" of scientific journals. *Journal of the American Society for Information Science and Technology*, 2007.
15. P. V. Marsden. Egocentric and sociocentric measures of network centrality. *Social networks (Soc. networks)*, 24:407–422, October 2002.
16. David R. Millen, Jonathan Feinberg, and Bernard Kerr. Dogear: Social bookmarking in the enterprise. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 111–120, New York, NY, USA, 2006. ACM.
17. S. Nanda and D. Kotz. Localized bridging centrality for distributed network analysis. In *Computer Communications and Networks, 2008. ICCCN '08. Proceedings of 17th International Conference on*, pages 1–6, 2008.
18. M. E. J. Newman. A measure of betweenness centrality based on random walks. September.