Combining Weighted Centrality and Network Clustering

Bohn, Angela; Theußl, Stefan; Feinerer, Ingo; Hornik, Kurt; Mair, Patrick; Walchhofer, Norbert

Published: 01/02/2009

Document Version
Publisher's PDF, also known as Version of record

Combining Weighted Centrality and Network Clustering

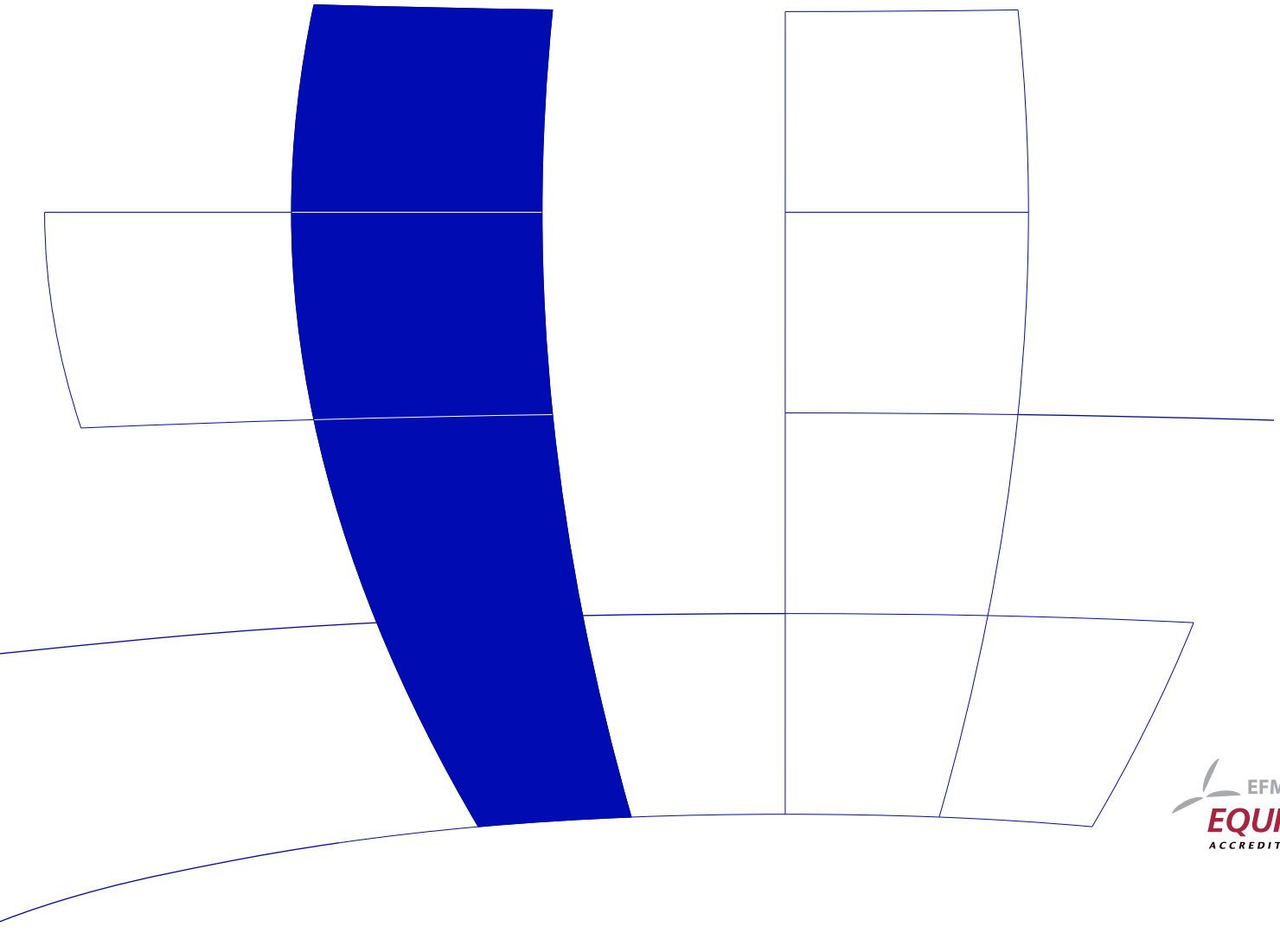
Angela Bohn, Stefan Theußl, Ingo Feinerer, Kurt Hornik, Patrick Mair, Norbert Walchhofer

Department of Statistics and Mathematics
WU Wirtschaftsuniversität Wien

Research Report Series

Report 97
December 2009

http://statmath.wu.ac.at/



Combining Weighted Centrality and Network Clustering

Angela Bohn, Stefan Theußl, Ingo Feinerer, Kurt Hornik, Patrick Mair, and Norbert Walchhofer

Abstract

In Social Network Analysis (SNA) centrality measures focus on activity (degree), information access (betweenness), distance to all the nodes (closeness), or popularity (pagerank). We introduce a new measure quantifying the distance of nodes to the network center. It is called *weighted distance to nearest center* (*WDNC*) and it is based on edge-weighted closeness (*EWC*), a weighted version of closeness. It combines elements of weighted centrality as well as clustering. The *WDNC* will be tested on two e-mail networks of the R community, one of the most important open source programs for statistical computing and graphics. We will find that there is a relationship between the *WDNC* and the formal organization of the R community.

1 Introduction

Until now, SNA centrality measures are based on the idea that a node should be considered more central if it is connected to a lot of other nodes or at least if its friends have many contacts. However, it depends on the question asked to a measure, if this interpretation of centrality makes sense. Imagine a president's wife who is maybe not very interested in politics and who has only a few contacts in a political network, but who has a large influence on her husband. Should she be considered central or not?

The *WDNC* is based on the idea that not only a node's integration into the network is important for its centrality, but also its distance to the center. In the scientific scene, not everyone feels the need to chat with dozens of people every day. However, such people may stay in contact with the network's information brokers, which guarantees him or her access to the most important news. The *WDNC* will be applied to the R [7] mailing lists R-help, designed to discuss users' questions, and R-devel, a communication platform for developers. We will find that the *WDNC* partly reflects the formal organization of the R community.

2 Methodology

The paper introduces a new measure called *WDNC*. Wasserman and Faust [8] provide an overview of the most frequently used centrality measures and clustering approaches. The WDNC is based on a widely used centrality measure called closeness [5]. It is defined as the normalized average distance (length of shortest path) from one vertex to all the others. A modification of closeness, the *EWC* [2], allows to take line values into account. It is defined as

$$EWC(i) = \frac{\sum_j \frac{\mathrm{llv}(i,j)}{d(i,j)}}{\max(\mathrm{lv})(n-1)}, \tag{1}$$

where $\mathrm{llv}(i,j)$ is the (average) last line value on the shortest path between i and j, $d(i,j)$ is the distance between i and j, $\max(\mathrm{lv})$ is the maximum line value in the entire network and n is the network size. The line value between i and j indicates the intensity of interaction and the distance between i and j is the length of the shortest path between them. The shortest path is the minimum number of edges needed to go from i to j. The last line value on the shortest path from i to j

		p			
		1	2	3	4
i	A	0.136	**0.145**	0.008	0.000
	B	**0.111**	0.111	0.008	0.000
	C	0.111	**0.151**	0.008	0.000
	D	0.049	**0.173**	0.012	0.000
	E	0.062	**0.167**	0.012	0.000
	F	**0.395**	0.019	0.000	0.000
	G	0.136	**0.142**	0.004	0.000
	H	0.012	0.062	**0.095**	0.003
	I	0.012	0.062	**0.095**	0.003
	J	0.012	0.049	**0.074**	0.006

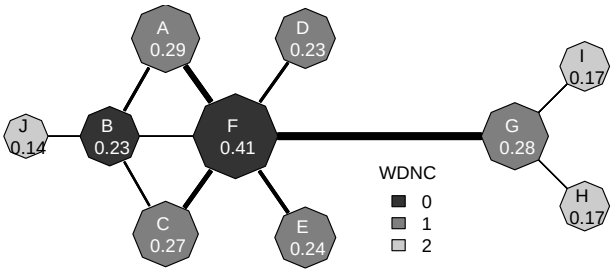


Figure 1: Example for the calculation of the *WDNC*

is then the line value between k and j, where k is the penultimate node lying on this path. k is identical with i if the distance between i and j is 1. The reason for considering only the last line value instead of using the sum or another aggregation of the line values is a matter of scaling. One could as well use the sum of line values on the shortest i-j-path. In this case, the larger the distance between i and j the more j's contribution to i's *EWC* is influenced by the line values between i and k. Taking only the last line values is more in line with the regular closeness, where each node contributes a certain distance and not a sum of distances. The impression that only the last line value is considered and the others are completely ignored is, however, false. When calculating i's *EWC*, all j are taken into account and thus all the lines lying on shortest paths contribute to i's *EWC*.

Splitting the sum in the enumerator into its summands and marking the distance in which a vertex gains the most *EWC*, corresponds to the definition of the *WDNC*. The *WDNC* of vertex i is defined as

$$WDNC(i) = \left(\inf \operatorname*{argmax}_{p} \sum_{j \in J_p(i)} \mathrm{llv}(i,j)/d(i,j) \right) - 1 \tag{2}$$

where $J_p(i)$ is the set of all nodes j which can be reached from vertex i by a path of length p. In words: The *WDNC* of a vertex i is the neighborhood p in which it gains the maximum *EWC* minus 1. If the maximum is not unique, infimum chooses the smallest p. The result may be interpreted as a line-weighted distance to the nearest center. The centers are vertices whose *WDNC* is 0. Thus, the *WDNC* combines elements of centrality measures, used to find influential nodes, and community detectors, serving to cluster nodes. (The R-code for the *WDNC* can be downloaded from http://R-Forge.R-project.org/projects/ewc/.)

Fig. 1 shows an example for the calculation of the *WDNC*. The matrix on the left shows the summands of the *EWC* enumerator with bold maxima. The corresponding graph with *WDNC* clusters in gray scale and *EWC* values as labels is on the right. The line strength symbolizes the size of the line values. The black vertices having an *EWC* of 0.23 and 0.41 form the center of the graph. Most of their neighbors' *WDNC* is 1. However, the node having an *EWC* of 0.14 has a *WDNC* of 2, because of its low adjacent line value. This shows that the *WDNC* does not calculate the distance to the nearest center, but the *weighted* distance to the nearest center: Vertices having high line values are closer to the center than nodes having low line values. It is important to notice that the vertices having a *WDNC* of 1 have higher *EWC* values than the black node with an *EWC* of 0.23. This illustrates that the *WDNC* is not the same as calculating *EWC* quantiles.

3 Data and Data Preparation

The characteristics of the *WDNC* are investigated using network data of the R-help and R-devel mailing lists during 2008. They serve to discuss questions from R developers and users, therefore they contain interesting information about a part of their social structure. Every e-mail sent to the mailing list is forwarded to all subscribers. They can be downloaded as compressed text files from https://stat.ethz.ch/pipermail/r-devel/ and

https://stat.ethz.ch/pipermail/r-help/, respectively.

Transforming Thread Trees to a Social Network Usually, mailing lists are represented as thread trees showing the referencing links between e-mails. Each e-mail has a message-ID and follow-ups additionally have reply-to IDs allowing to build thread trees [4]. The next data preparation step consisted in transforming the thread trees, where nodes represent e-mails, in such a way that nodes represent e-mail authors. We drew an edge between an author and his or her "forefathers" in the thread tree. The networks are represented as weighted matrices, where the weights correspond to the number of e-mails exchanged between two authors. To calculate the *WDNC*, we need the networks to be strongly connected. As the largest strongly connected subgraph (component) cover only a small part of the network members we symmetrized the networks using the sum of incoming and outgoing arc weights (sum of sent e-mails and received e-mails) and only the largest component was considered. The other components (42 in R-help and 15 in R-devel) have only one to three members and are therefore negligible.

Finding Aliases The second data preparation step consisted in finding aliases, as authors may have several different user names and e-mail addresses. Like Bird et al [1] we first normalized the user names and e-mail addresses, then we used the Levenshtein distance [6] to find clusters of similar names. To increase the probability of finding all aliases, we allowed a distance of 0.3 between the names within one cluster. Thus, each cluster contained a number of strings that differed in at most 3/10 of the symbols. We checked those clusters manually and rejected 60% of them, so we expect to have found most aliases. This way, the R-help network was reduced from 5128 to 4065 nodes and the R-devel network from 837 to 652.

Description of R-Help Network The largest component of the network has 3672 nodes, its diameter (length of longest shortest path) is 7, the average degree (number of direct neighbors) is 11.8 and the median degree is 4. Each network member wrote 7.6 e-mails on average. The maximum of e-mails sent was 1071 by Brian Ripley. 1640 people wrote only one e-mail. 76% of the line values (number of e-mails exchanged between two authors) is 1. The maximum of e-mails exchanged between two authors was 72 (Gabor Grothendieck and Brian Ripley) and their mean is 1.5.

Description of R-Devel Network The largest component of the R-devel network has 566 nodes. Its diameter is 6. As the network is much smaller, the average degree is 8.5, but the median degree is also 4. Brian Ripley is by far the most active author in the R-devel network. He sent 522 e-mails and his degree is 332. The second most active author, Duncan Murdoch, wrote only 255 e-mails and his degree is 177. Most line values (67%) are 1 and the their mean is 1.9. The maximum of e-mails exchanged between two authors was 63 (Brian Ripley and Duncan Murdoch).

4 Results

In this section, we will apply the *WDNC* presented in Sect. 2 to the networks described in Sect. 3. The results will be compared to other centrality measures. Finally, the informal structure of software development will be compared to the formal organization.

4.1 R-Devel Network

In the R-devel network, we identified five very central authors using the *WDNC*: Peter Dalgaard, Gabor Grothendieck (GG), Martin Mächler, Duncan Murdoch, and Brian Ripley (BR). They are in close contact to each other, so the network does not to have several separated centers with each having its own community, but it is monocentric. BR is the most active author in terms of degree and number of e-mails sent. Many vertices adjacent to BR do not have any other contacts. He prevents the network from being split into many small components. In contrast, some of BR's

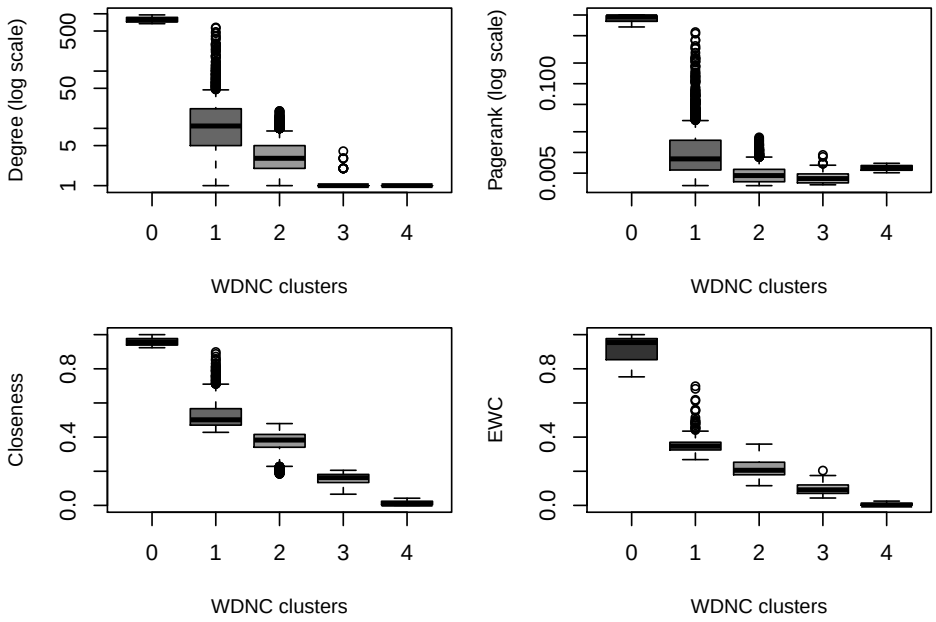


Figure 2: Boxplots of *WDNC* vs. centrality measures in the R-help network

neighbors are more active and they are well connected to other network members, so they may be considered to be the core of the network. 69% of the network members have a *WDNC* of 1. Most of them are neighbors of the central cluster. 28% have a *WDNC* of 2 and most of them are neighbors of those having a *WDNC* of 1.

4.2 R-Help Network

In the R-help network we identified three very central authors: GG, Jim Holtman, and BR. Like in the R-devel network, the central nodes are in close contact to each other. However, in this network, each of the vertices having a WDNC of 0 have a community that is partly separated from the others. Furthermore, all nodes in the central cluster are comparably active. These observations indicate, that BR's position is not as marked as in the R-devel network.

4.3 Empirical Evidence of the Usefulness of the *WDNC*

As the *WDNC* defines a vertex' importance according to its weighted distance to the nearest center, it is crucial to know whether the choice of centers is reasonable. Fig. 2 shows boxplots of centrality measures for each *WDNC* cluster (x-axis) in the R-help network. The vertices having a *WDNC* of 0 are far more central than the other clusters according to degree and pagerank [3]. Compared to closeness and *EWC*, the difference between cluster 0 and the others is smaller, because the *WDNC* is based on these measures. The corresponding boxplots of the R-devel network are very similar (Fig. 3), so we conclude that in the mailing list networks, the algorithm chose a few very central vertices to have a *WDNC* of 0. Nodes with a *WDNC* of 1 are clearly less central, however, the amount of outliers in this cluster (139 to 140 in R-help and 23 to 25 in R-devel) indicates that it is heterogeneous.

4.4 *WDNC* Compared to Formal R Organization

The formal organization of R is not as strictly defined as in software companies. However, the R community can be roughly divided into several groups of developers and users. There are

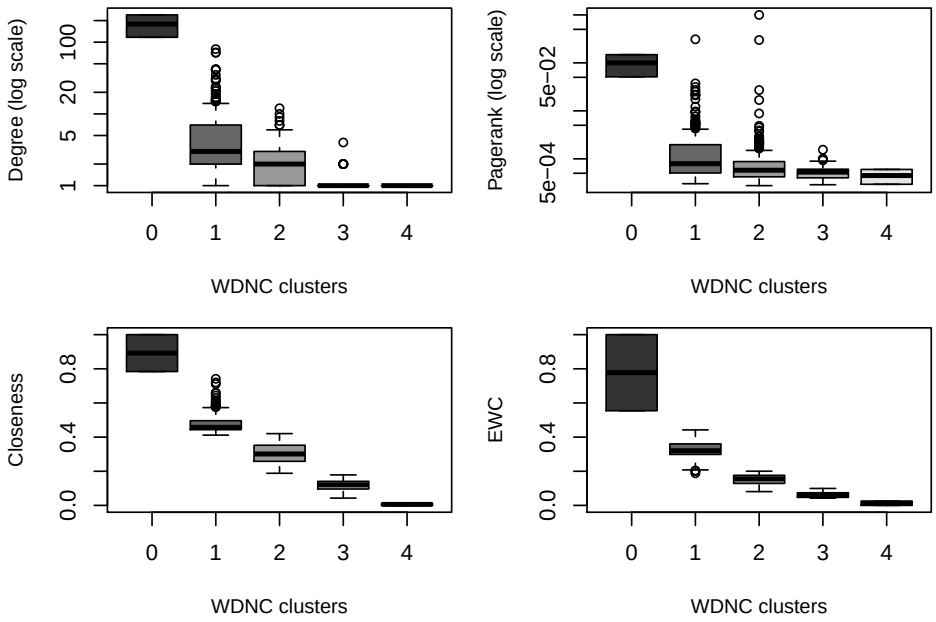


Figure 3: Boxplots of *WDNC* vs. centrality measures in the R-devel network

Table 1: Cross-classified table of *WDNC* vs. developer and user groups

R-devel						R-help					
WDNC	c.d.[a]	m.d.[b]	o.d.[c]	users	sum	*WDNC*	c.d.[a]	m.d.[b]	o.d.[c]	users	sum
0	4	0	1	0	5	0	1	0	1	1	3
1	11	11	186	185	393	1	11	13	411	961	1396
2	0	2	64	94	160	2	3	3	581	1552	2139
3	0	0	3	5	8	3	0	2	33	96	131
4	0	0	0	0	0	4	0	0	1	2	3
sum	15	13	254	284	566	sum	15	13	1027	2612	3672
mean	**0.7**	**1.2**	**1.3**	**1.4**	1.3	mean	**1.1**	**1.4**	**1.6**	**1.7**	1.7

[a] core developers [b] main developers [c] other developers

19 core developers and 44 main developers who are mentioned on the R Core Team website (http://www.R-project.org/contributors.html). In addition, there are hundreds of other developers whose names can be obtained from the R package descriptions on CRAN (http://CRAN.R-project.org/). The group of users can be divided into active and passive users. Active users report bugs, make suggestions for improvements and write to the mailing lists. Passive users do not communicate their experiences, but only use the software. Thus, the mailing lists contain only information about active users. However, we cannot distinguish between the different kinds of active users. Table 1 shows a cross-classified table of *WDNC* vs. developer and user groups. It shows that, although R-devel is inteded for developers and R-help for users, a separation between users and developers is only partly realized: Half of the R-devel authors are users and 29% of the R-help authors are developers. (Note that some developers might be classified as users if their package is not yet on CRAN.) However, inside the networks, the behavior of the groups differs. If we take the membership of an author to a certain developer or user group as an indicator for the level of commitment of this author to R, where the membership to the core developers corresponds to highest commitment and the membership to the user group means lowest commitment, we see that the mailing list behavior reflects these differences: The core developers have the lowest average *WDNC* in both networks (0.7 and 1.1), which means that they are most central. The group of main developers is slightly less central (1.2 and 1.4) and the other developers have an average *WDNC* of 1.3 and 1.6. Finally, many users are located at the periphery, which results in an average *WDNC* of 1.4 and 1.7. (Like any other centrality measure, the *WDNC* of a vertex

can only be interpreted in comparison to nodes of the same network and not across networks: An average *WDNC* of 1.4 can indicate a central position in one network and a peripheral position in another.) Thus, a low *WDNC* is associated with high commitment.

5 Conclusion and Discussion

This paper introduced a combination of centrality measure and clustering approach called *WDNC*. The *WDNC* was applied to the OSS mailing lists R-devel and R-help. We found that the network structure of both mailing lists are similar: They are monocentric and dominated by a few very active e-mail authors staying in close contact to each other. This can be explained by the fact that the mailing lists do not reflect a stringent separation between developers and users. However, the *WDNC* reveals that the behavior of users and types of developers differs. If we take a developer's formal role as indicator for his or her commitment to R, where the membership to the core development group indicates highest commitment and being a user indicates lowest commitment, we see that a low *WDNC* is associated with high commitment. Thus, the level of commitment tends to be reflected by a central and influential position in the mailing lists. However, the validity of the results is restricted to the communication via mailing lists which capture only a small part of the social behavior. Although the data structure did not allow to use the directed version of *WDNC*, it can be useful in other applications, for example to distinguish question-people from answer-people.

References

[1] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan. Mining email social networks. In *Proceedings of the 2006 International Workshop on Mining Software Repositories, Shanghai, China*, pages 137–143, 2006.

[2] A. Bohn, N. Walchhofer, P. Mair, and K. Hornik. Social network analysis of weighted telecommunications graphs. Report 84, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series, 2009. URL http://epub.wu.ac.at/.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[4] I. Feinerer, K. Hornik, and D. Meyer. Text mining infrastructure in R. *Journal of Statistical Software*, 25(5):1–54, 2008. URL http://www.jstatsoft.org/v25/i05/.

[5] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1:215–239, 1979.

[6] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8, Soviet Physics Doklady, 1966.

[7] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL http://www.R-project.org/.

[8] S. Wasserman and K. Faust. *Social Network Analysis – Methods and Applications*. Cambridge University Press, 1997.