



WeKnowIt

Emerging, Collective Intelligence for Personal,
Organisational and Social Use

FP7-215453

D4.1.2

Community Analysis Tool and Evaluation

Dissemination level	Public
Contractual date of delivery	30.09.09
Actual date of delivery	08.10.09
Workpackage	WP4 Social Intelligence
Task	T4.3 Community Analysis Tool
Type	Prototype
Approval Status	Draft
Version	1
Number of pages	22
Filename	D4.1.2-man_2009-9-30_v10_emka_deliverable-community_analysis_tool_public_version-1.odt

Abstract

The Community Analysis Tool provides a package of indices, which allow a fast, scalable and rational analysis of the community structures. Algorithms have been designed and implemented to assure this goal. In addition to the standard measures already implemented in D4.1.1 as an innovative measure the complex valued centrality measure is included.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.



co-funded by the European Union

History

Version	Date	Reason	Revised by
1	28.08.09	Creation / Structure	Andreas Sonnenbichler
2	02.09.09	Method and Applications	Bettina Hoser
3	03.09.09		Michael Ovelgönne
5	14.09.09		Andreas Sonnenbichler
6	17.09.09		Bettina Hoser
7	22.09.09	Incorporate Review comments	Michael Ovelgönne, Andreas Sonnenbichler
8 to 10	08.10.09	Consortium internal review comments included (reviewer: Pavel Smrz - BUT)	Bettina Hoser

Author list

Organization	Name	Contact Information
EMKA	Michael Ovelgönne	Michael.Ovelgoenne@kit.edu
EMKA	Bettina Hoser	Bettina.Hoser@kit.edu
EMKA	Andreas Sonnenbichler	Andreas.Sonnenbichler@kit.edu

Executive Summary

The Community Analysis Tool (CAT) provides a package of social networking indices. Social Networking Analysis (SNA) provides methods to analyse networks. Networks can be e.g. friendship networks but also any other data which can be provided as network.

The CAT implements algorithms, which allow a fast, scalable and rational analysis of the community structures. In addition to the standard measures already implemented in D4.1.1 as an innovative measures the complex valued centrality measure is included.

Being mainly a back-end service within the WeKnowIt service landscape, CAT can analyse any network for other modules. E.g. friendship networks can be analysed concerning clusters. Communication networks within an emergency case can be monitored by the complex Eigenvector centrality to detect unusual communication patterns.

Accordingly to the WeKnowIt project plan, this deliverable focuses on the methods and algorithms of the Community Analysis Tool. Applications and usage in other services will be mentioned, but is not in the focus of this document.

Abbreviations and Acronyms

CAT	Community Analysis Tool
CAP	Community Administration Platform
SNA	Social Network Analysis
iCAT	Initial Community Analysis Tool

Table of Contents

1.Introduction.....6
 1.1.Description of the extension into the CAT.....6
2.Related Work and Methodology.....9
3.Implementation.....14
4.Application.....15
 4.1.Emergency Response Use Case.....15
 4.2.Consumer Group Use Case.....16
 4.3.Community Administration Platform.....17
5.Evaluation.....18
6.Conclusion.....21

Illustration Index

Index of Tables

1. Introduction

The community analysis tool (CAT) is a library of methods designed to provide social network analysis both for the other work packages and for end users. It is a modular set of classes based on the JUNG framework and designed to be flexible and integrable in various application contexts.

1.1. Description of the extension into the CAT

This deliverable D4.1.2 continues the work described in D4.1.1 and completes the initial community analysis tool (iCAT) to the full community analysis. The iCAT consists of data structures to store graphs and basic methods to analyse them. Furthermore, the iCAT covered a database cache for graphs and analysis results. This extension of the community analysis tool provides advanced analysis functionality.

The iCAT provided analysis methods on three levels: actor, sub-network and network level. Actor level methods look at all actors separately and assign values to them e.g. to rank them by their prestige or their influence. Sub-network level methods look at groups of actors and decompose the network into sub-networks of cohesive actors or sets of actors with similar attribute properties. Finally, network level methods examine the network as a whole and provide results that describe the entire network.

The iCAT provided beside the local measure degree centrality three further actor level algorithms that are based on shortest-paths: closeness centrality, inverse distance closeness centrality and betweenness centrality. We introduce into our advanced library the analysis of the eigensystem of graphs. Instead of the conventional eigenvector centrality our complex eigensystem approach reveals for directed networks much more information on the network structure. This complex eigensystem approach is especially valuable for the analysis of directed networks such as email networks or telephone networks.

Figure 2 shows an UML diagram of all actor level analysis methods which have a very light-weight interface. The layout of the required graph data structures is shown in Figure 1. The architecture of the community analysis tools remains unchanged. For further details we refer to deliverable D4.1.1.

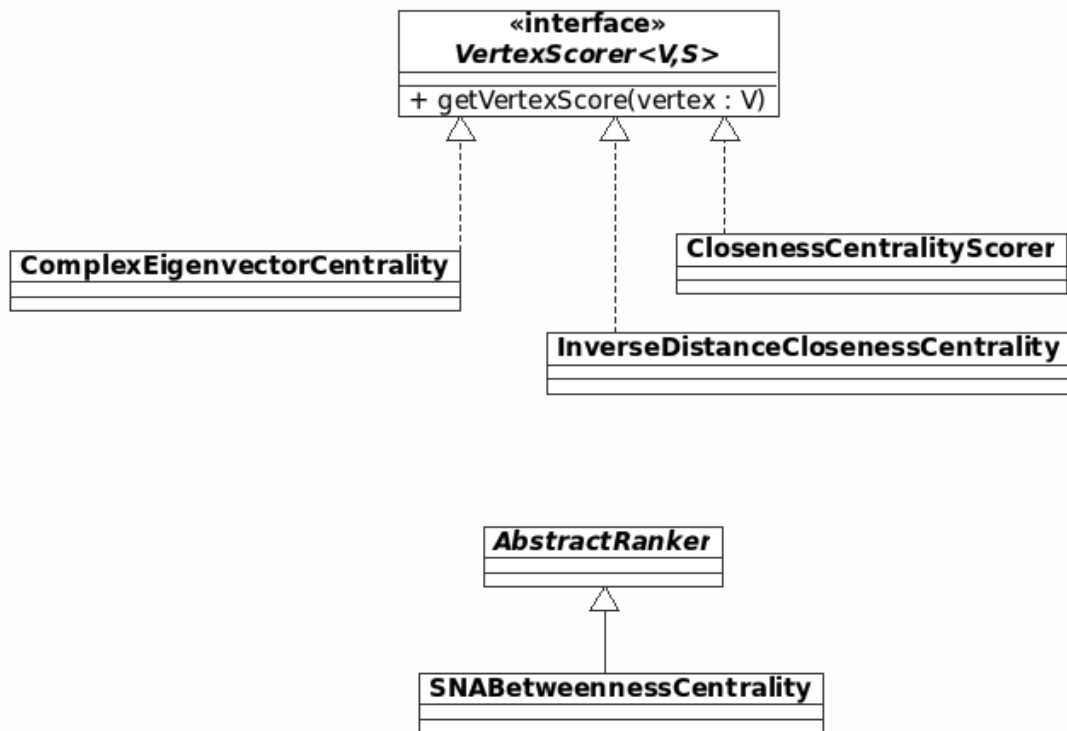


Figure 2: UML class diagram of the actor level analysis methods

2. Related Work and Methodology

One of the major problems when analysing real world communication networks is, that, more often than not, they are asymmetric. This means that the communication of actor A in the direction of actor B is not of the same amount or strength as that in the reverse direction.

To calculate the eigenvector centrality, a standard measure for the centrality of an actor in a network, for such asymmetric networks the standard procedure is to symmetrize the resulting adjacency matrix. This is necessary since the eigensystem calculation for asymmetric matrices leads into the following difficulties:

- 1) Eigenvalues may become complex valued. As of now no meaningful interpretation has been proposed for such complex valued eigenvalues in SNA.
- 2) Eigenvalues and/or eigenvectors may be degenerated. This means that non-zero eigenvalues are not simple, and thus more than one eigenvector for each eigenvalue can be found. For this, other than in e.g. physics, no meaningful interpretation is available in SNA.

Thus real asymmetric adjacency matrices are transformed into symmetric matrices, since these have the following characteristics:

- 1) All eigenvalues are real and simple.
- 2) The eigenvectors build a complete orthogonal eigenspace.

There are different ways to symmetrize any real valued square matrix A (adjacency matrix of the observed network):

- 1) Take the symmetric part of $A = 1/2 (A + A^T) + 1/2 (A - A^T)$
- 2) Multiply either from left or right with the transpose of the original matrix. $B = AA^T$ or $C = A^T A$.
- 3) Any other form of symmetrization, like making the matrix binary and set symmetric entries to 1 if any of the two entries is 1, and 0 else.

A major drawback of this is that in the process of symmetrization information is lost or neglected. As Fagiolo (2006) showed, this loss of information becomes more and more distorting as the size and structure of a network changes. This can be shown by the use of a distance-to-symmetry measure. If a network can be regarded as close to symmetry, one may use standard indices of centrality. But when the network becomes more and more asymmetrical, the results of a symmetrization may be very misleading.

There are some approaches to calculate the eigenvector centrality for real asymmetric matrices. Bonacich and Lloyd (2001) presented an approach

for directed networks. But they only consider one directed link between any two actors. This is a restriction in real world networks and practically would amount to a dichotomization in which if the communication from actor K to actor L is 'larger' than that in the reverse direction, the adjacency matrix entry would be 1 in entry a_{kl} and 0 in a_{lk} .

To take in all available information Hoser and Geyer-Schulz (2005) presented another approach, which is implemented in this deliverable.

This approach is based on the possibility to represent the inbound and outbound links by just one number, which still retains the distinctiveness of the information without information loss. This can be achieved by using complex numbers. A complex number z can be represented in two different ways:

$$z = p + iq = re^{i\varphi}$$

with p the real part of z ($\text{Re}(z)$), q the imaginary part of z ($\text{Im}(z)$), r the absolute value of z ($\text{abs}(z)$), φ the argument or phase of z ($\text{arg}(z)$), and i the imaginary unit ($i^2 = -1$).

This approach uses a transformation of the real valued weighted adjacency matrix A (especially if asymmetric) with all entries on the diagonal equal to zero into a complex hermitian matrix H .

$$H = (A + iA^T) e^{-i\pi/4}$$

with A^T the transpose of A .

This matrix is hermitian. The mathematical characteristics of such matrices are that all eigenvalues are real. The eigenvectors form an orthogonal eigenbasis. From this it also is clear that

$$H = \sum_j \lambda_j P_j$$

with $P_j = x_j^* x_j$ the orthogonal projector and x_j^* the complex conjugate transpose of the column vector x . Thus H can be understood as the weighted sum of projectors.

This sum, called a Fourier-sum, allows another interpretation of each eigenvector. Now each eigenvector is a signal, which, weighted with its eigenvalue, and added to all existing signals gives the complete 'communication' signal. The eigenvector with the largest positive eigenvalue thus represents the strongest communication pattern in the analysed network.

We use this transformation to analyse two-way-asymmetric communication in the following way.

First we can analyse the spectrum (set of all non-zero eigenvalues). If the spectrum shows a complete symmetry (the largest positive eigenvalue = - the largest negative eigenvalue, and so forth), then the communication pattern is most probably dominated by star-like structures as seen in Fig. 3.

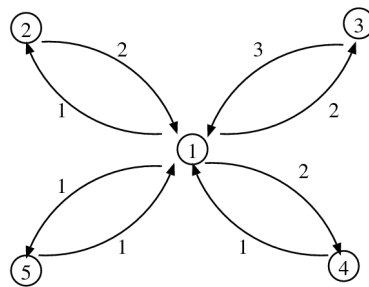


Figure 3: Star-like structure

The corresponding eigensystem for the non-zero eigenvalues is given in Table 6.

Eigenvalues	0.5		-0.5	
Eigenvectors	abs	arg	abs	arg
x1	0,71		0,71	
x2	0,32	2,84	0,32	-0,32
x3	0,51	0,51	0,51	-0,2
x4	0,32	-2,82	0,32	0,32
x5	0,2	$\pi\pi$	0,2	0

Table 6: Eigensystem of star-like structure

If, on the other hand the spectrum has just one positive eigenvalue and all other non-zero eigenvalues are negative and of the same value, then the communication pattern is that of a complete graph.

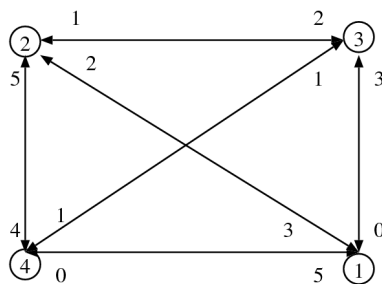


Figure 4: Complete graph

Eigen-- values	1		-0,67		-0,37		0,05	
Eigen- vectors	abs	arg	abs	arg	abs	arg	abs	arg
x1	0,51	0,53	0,41	-1,9	0,64	0	0,4	-1,7
x2	0,56	0,21	0,53	3,1	0,59	2,2	0,23	1,4
x3	0,29	0,75	0,2	1,9	0,48	-1,7	0,8	0
x4	0,58	0	0,72	0	0,1	-2,3	0,37	2

Table 2: Eigensystem of Fig. 4

One can show how the spectrum departs from either of these two extremes when perturbations arise. This can be used for example to see how a structured communication (e.g. emergency scenario) becomes unstructured under pressure, or how structures emerge from otherwise amorphous link structures.

The second step is the analysis of the eigenvectors. The absolute value of the eigenvector component is similar to the standard eigenvector centrality measure of SNA. But in addition, since the eigenvectors are complex valued, each eigenvector component now also has a phase. This phase gives information about the relative direction of communication between any two actors within the given network, as can be seen by the transformation for the hermitian matrix.

If the network structure is e.g. a star structure as in Fig.3 of just one star it can be shown that, as described above, the spectrum only has two non-zero eigenvalues of the same absolute value but different sign. The corresponding eigensystem reveals the star structure in that the centre of the star has the largest absolute value and the 'rays' of the star have the same absolute value in both eigenvectors but show a phase shift of π between the two vectors. While within each vector the phase is such that the clear directionality of all actors can be seen by using the phase of each component. In the other extreme case of a complete graph, the eigenvector the largest (positive) eigenvalue shows the complete graph structure, while the remaining eigenvectors to the non-zero eigenvalues are mathematically a basis for the remaining communication patterns.

The shift between the extreme patterns can be shown by using perturbation theory, e.g. Kato (1995). There is extensive literature on perturbation of hermitian operators (e.g. Ipsen 2003, Dopico et al. 2000).

This method can also be used for spectral clustering, as was shown by Hoser and Bierhance (2007).

In applications this approach has already been used to look into organizational efficiency (Hoser and Geyer-Schulz 2007).

3. Implementation

- classified -

4. Application

The application in the use cases are given below.

In this context it might be of interest to detect misuse or cheating of centrality in a network. This can only be done when a hypothesis on user behaviour or a clear communication protocol are ex ante known about the community. This could be the case in the emergency scenarios. If such an ex ante knowledge would be available the network structure can be derived from it and can be checked against the actual structure. This test should be performed by the administrator, since in some instances small differences might be relevant, in others only abrupt changes need to be handled.

4.1. Emergency Response Use Case

In an emergency there are several different kinds of communication that arise.

- 'Official structured' communication. This includes all communication between official organisations like city council, police, hospitals, government, etc. This is a highly structured and regulated kind of communication. It is clear who communicates with whom and how in case of emergency. Thus this network is most probably asymmetric to reduce load on networks, well defined and structured.
- 'Official unstructured' communication. In times of crisis communication may become unstructured. In addition, not all communication is regulated in an organization. This may happen under stress. A communication pattern that is supposed to be structured and suddenly becomes unstructured is a sign of stress. Thus an early warning about the start of such a destructuralization might enhance efficiency of official communication.
- 'User' communication. This communication is unstructured at best. During crisis it may even become chaotic. One of the major problems arising from this is the breakdown of infrastructure like telephone or mobile communication lines. Post crisis this still goes on a high level, until everyone who is trying to find information has found the communication partner he needed or gives up. This communication is highly asymmetrical.

The analysis of asymmetric communication in an emergency case can thus provide information and warning signs. As described above the method implemented here can help to detect a shift in communication behaviour. Thus the method can be used to give an early warning to officials at e.g. a city council that the official structured communication is starting to become unstructured, in which case immediately counter measures should be taken.

An analysis is planned to follow these steps:

- 1) Ex ante: The eigensystem of the official structured communication pattern has to be available
- 2) First signs of a crisis appear. The network load rises. The eigenvalues will show this by rising. As long as the eigensystem of the resulting communication structure still adheres to the protocol designed for that case by the emergency organisations, the structure of the eigenvectors will not change.
- 3) When the structure of the spectrum starts to shift, e.g. shows a tendency towards clique building or star patterns or on the other hand shows a tendency to become amorphous, the eigenvalues will give a first hint. If a threshold has been defined for the change of the spectrum, a first warning note could be generated. Then a look at the perturbation of the eigenvectors would need to follow.
- 4) If the perturbation starts to change the eigenvectors, thus starts to change the structure of the communication, one could again define thresholds to make changes visible which would be indicators of communication breakdown. If e.g. communication should be very structured, but the eigenvectors now start to show a less and less structured communication, e.g. the eigensystem of a complete graph, then a warning should be triggered to make the responsible person aware of the pending problem. It would also be the responsibility of the organization to define intervention procedures for communication shifts of that kind.

On the user side it would be possible to allow for a more structured communication like proposed in one of the services of the project, namely the Emergency Notification Service.

4.2. Consumer Group Use Case

In the Consumer Group Case our approach can be used to support the administration of a group. This would be best achieved by calculating the eigensystem in a dynamical way. The time steps should be adapted to the 'rhythm' of the observed communication. Thus it could be used by the Community Administration Platform (CAP).

In an anecdotal (due to the fact that the data set was already given, statistical validity or the possibility to generalize are not given) analysis performed as a master thesis on the basis of the LycosIQ data set we found, that groups that centre around questions and answers show a very different communication behaviour than groups where constant interaction takes place. In addition we could show how the use of a guest account distorts communication patterns (Schaefer and Hoser, 2008).

The analysis of a consumer group communication could follow these steps:

- 1) Define what kind of communication structure the proposed consumer group most likely will have. E.g. a question and answer group will show a highly directed graph structure - depending on the modelling of the network. In addition, when the time evolution is analysed the interactions between any two actors will be short lived. On the other hand if it is a friendship network the structure of the communication will most probably be very homogeneous. The corresponding eigensystems will either look like stars (question and answer) or like complete graphs (friendship). In addition the distance to symmetry could be used to see how balanced the communication is.
- 2) The administrator of a given group could thus use the eigensystem to see whether the group is active in the defined sense or whether the communication pattern changes. The eigenvalues show the communication volume. The eigenvectors show the structure.
- 3) If the structure starts to deviate from the expected structure, the administrator could intervene. Thus e.g. if the community seems to die, which would mean decrease of volume and/or visible shift to a more or less dyadic or a more amorphous structure (depending on the original structure), the administrator could try to boost activity by introducing new people or new topics. Or he could shut down this group and recommend to join other groups to the remaining users.

4.3. Community Administration Platform

- classified -

5. Evaluation

To evaluate the presented method we could not rely on the only community data set (LycosIQ) made available to WeKnowIt as this dataset has never been systematically analysed and thus, there is no standard of comparison.

To evaluate our approach we have used a very well investigated data set, the EIES network by Freeman (Wasserman and Faust, 1999). This network is asymmetric as can be seen in Table 3.

	A	B	C	D	E	F	G
A	0	115	17	93	53	33	84
B	84	0	4	5	5	0	15
C	16	10	0	15	3	3	4
D	127	22	17	0	57	12	34
E	57	9	4	57	0	8	10
F	23	4	3	9	8	0	33
G	118	24	5	35	15	45	0

Table 3: EIES Data set

At first sight this looks like a complete graph. A look at the eigensystem will show more detail, e.g. a structure within this graph.

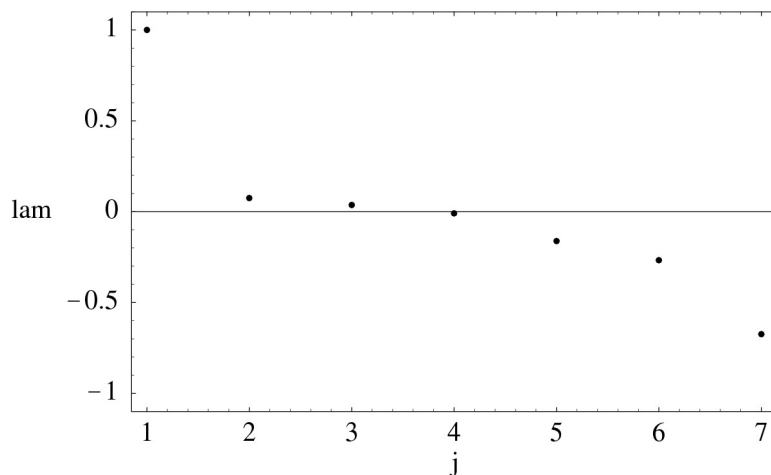


Figure 5: Eigenvalues of EIES data set

As can be seen there is the suggestion of a symmetry for the largest and smallest eigenvalue (λ -axis), which could mean that there is a star-like structure hidden in the data set. But since the symmetry is not complete it looks like a perturbed star or, coming from our earlier observation, a perturbed complete graph.

The eigensystem of this data set is given in Table 4.

EVa	1,00		-0,67		-0,27		-0,16		0,07		0,04		-0,01	
EVe	abs	arg	abs	arg	abs	arg	abs	arg	abs	arg	abs	arg	abs	arg
x1	0,62	0,00	0,76	0,00	0,06	-0,09	0,07	-2,86	0,05	-0,96	0,15	-0,02	0,06	-2,73
x2	0,34	0,18	0,39	-3,00	0,06	-1,88	0,44	0,15	0,20	-0,92	0,69	0,00	0,12	-2,63
x3	0,10	-0,09	0,02	-2,32	0,12	3,12	0,02	0,29	0,16	2,93	0,12	-0,30	0,97	0,00
x4	0,45	-0,15	0,38	3,00	0,65	0,00	0,21	2,99	0,38	2,75	0,20	2,64	0,02	1,67
x5	0,29	-0,07	0,08	-3,00	0,56	3,10	0,39	-0,04	0,53	2,64	0,38	2,56	0,14	3,03
x6	0,17	0,05	0,03	-1,97	0,22	0,02	0,55	0,00	0,56	0,00	0,52	-2,94	0,16	0,34
x7	0,41	-0,15	0,34	2,93	0,45	3,07	0,55	2,91	0,44	-0,25	0,16	-2,89	0,01	1,70

Table 4: Eigensystem for EIES data set

As can be seen there is indeed a star-like pattern visible. Actor A is the center of this star and actors B, D and F are connected to him. This can be seen by the phase shift of approximately 3,14 between the arg of the same actor in the two eigenvectors. The center of the star has the largest abs-value. The three 'rays' are the next strongest.

To compare this result with standard procedure results we also show what happens to the eigensystem if we use the transformation AA^T . This provides a real symmetric matrix with real valued eigensystem. This transformation puts the emphasis on the outbound behaviour similarities between any two actors. Thus the eigensystems reflects whether two actors had a similar outbound behaviour to all other actors in the network.

What can be seen immediately in the table below is the different behaviour of the eigenvalues. They drop off far more rapidly betweenness the second and third eigenvalue, thus pointing towards a star like structure. While in the previous calculation the star is visible but well embedded into a complete graph structure. In addition the structures in the third and fourth eigenvector are not highly relevant, as is seen by the low corresponding eigenvalue.

EVa	1,00	0,44	0,07	0,02	0,01	~0,00	~0,00
EVe							
x1	-0,52	-0,83	0,07	0,07	-0,05	-0,13	0,09
x2	-0,28	0,30	0,04	-0,39	-0,34	-0,72	0,20
x3	-0,1	-0,01	-0,12	-0,05	0,12	-0,27	-0,94
x4	-0,54	0,34	0,62	0,07	0,42	0,15	-0,02
x5	-0,29	0,04	-0,58	-0,54	0,48	0,20	0,13
x6	-0,14	-0,02	0,19	-0,50	-0,58	0,55	-0,22
x7	-0,50	0,30	-0,46	0,55	-0,35	0,17	-0,02

As a conclusion we can state that if we assume that a network shows different communication patterns, as is a well founded assumption in real life communication networks, a loss of information by a standard transformation might lead to incorrect results. This might be especially harmful if actions within an emergency use case are being planned on the basis of a communication network analysis.

6. Conclusion

The proposed complex valued eigenvector centrality is a measure for centrality in asymmetric networks. This becomes even more visible when the network acquires a larger size, as we can assume it will in any of the two use cases.

The method is innovative as it has not yet been implemented in any other social network analysis framework due to extensive test in different applications such as mobile communication (master thesis based on the Reality Mining data set of the MIT Media Lab), forecasting stock markets, or controller networks (in a project with industry partners). In some of these projects already the possibility to detect intended misuse has been analysed (forthcoming: PhD thesis of Jan Schröder; KIT).

Thus we see a high potential to help administrators of communities to better keep track of the 'life' of their communities and to be able to intervene if either decline of the community or misuse must be assumed.

References

- P. Bonacich and P. Lloyd, 2001, *Eigenvector-like Measures of Centrality for Asymmetric Relations*, Social Networks, 23, p.191-201
- C. Schaefer, B. Hoser, 2008, *Die Beeinflussung von Zentralitätsmaßen der sozialen Netzwerkanalyse durch Gästeaccounts in Internet-Diskussionsforen (in german)*, in: Netzwerkanalyse und Netzwerktheorie, ed. Ch. Stegbauer, VS Verlag, Wiesbaden, p.273-286
- F. Dopico, J. Moro and J. Molera, 2000, *Weyl-type relative perturbation bounds for eigensystems of Hermitian matrices*, Linear Algebra and its Applications 309 (1), p.3 - 18; G. Stewart, J. Sun, 1990, Matrix Perturbation Theory, Academic Press, Inc, London
- G. Fagiolo, 2006, *Directed or Undirected? A New Index to Check for Directionality of Relations in Socio-Economic Networks*, Economics Bulletin 3 (34), p.1-12
- B. Hoser and Th. Bierhance, 2007, *Finding Cliques in directed weighted graphs using complex hermitian adjacency matrices*, in: Proceedings of the 30th Annual Conference of the German Classification Society, eds. R. Decker and H.-J. Lenz, Advances in Data Analysis, Springer, p. 83 - 91
- B. Hoser and A. Geyer-Schulz, 2007, *Organisationseffizienz (in german)*, in: Analyse sozialer Netzwerke und Social Software - Grundlagen und Anwendungsbeispiele, eds. C. Müller, N.Gronau, GITO Verlag Berlin, p.133-155
- B. Hoser and A. Geyer-Schulz, 2005, *Eigenspectralanalysis of Hermitian Adjacency Matrices for the Analysis of Group Substructures*, Journal of Mathematical Sociology, 29 (4), p.265-294
- I. Ipsen, 2003, *A note on unifying absolute and relative perturbation bounds*, Linear Algebra and its Applications 358, p.239-253
- T. Kato, 1995, *Perturbation Theory for Linear Operators*, 2nd edition, Springer, New York
- S. Wasserman, K. Faust, 1999, *Social Network Analysis and its Applications*, Cambridge University Press