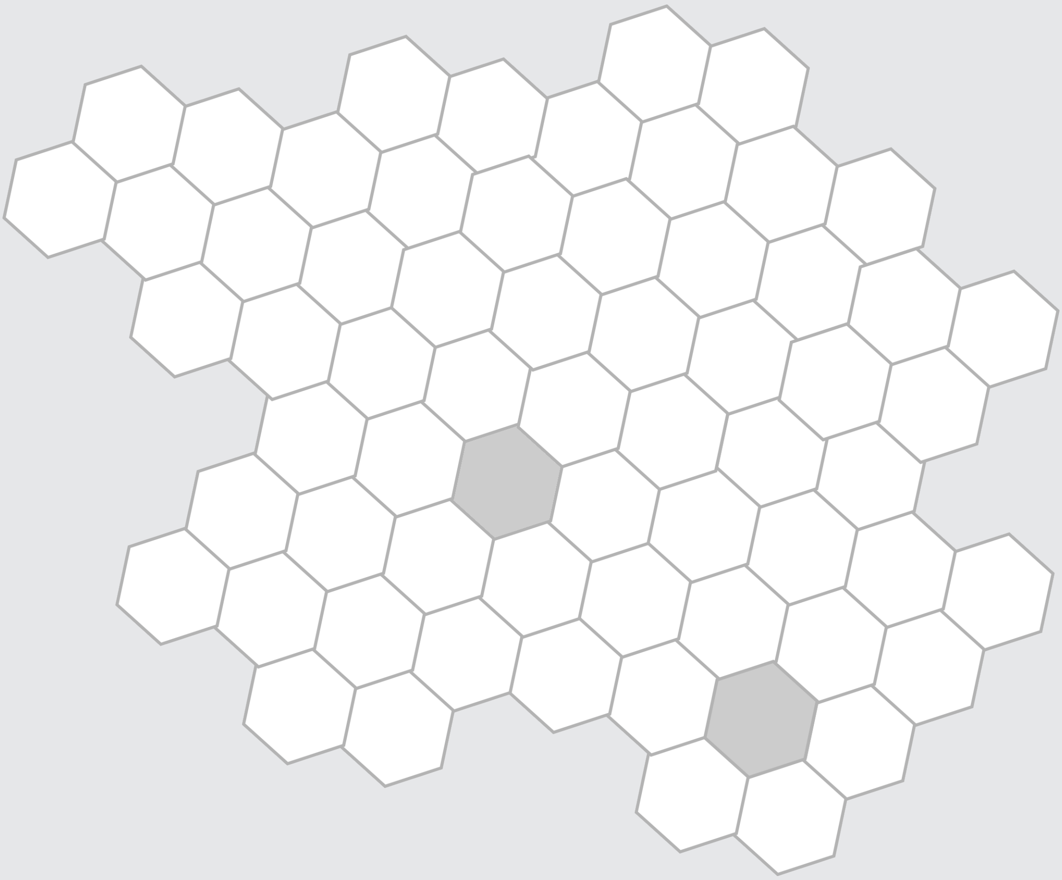
International Journal on SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining



II, 2011, 1–2
ISSN 1804-5251

...OFFPRINT...

International Journal on
SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining

II, 2011, 1–2

Editor-in-Chief: Jan Žižka

Publisher's website: www.konvoj.cz
E-mail: konvoj@konvoj.cz, SoNet.RC@gmail.com
ISSN 1804-5251

No part of this publication may be reproduced, stored or transmitted in any material form or by any means (including electronic, mechanical, photocopying, recording or otherwise) without the prior written permission of the publisher, except in accordance with the Czech legal provisions.

Social Web Mapping Dynamic Analysis of Social Networks

Kristína Machová
Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University in Košice, Letná 9, 042 00 Košice, Slovakia
Tomáš Rakušinec
Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University in Košice, Letná 9, 042 00 Košice, Slovakia

Abstract

This paper deals with social networks, their types, and an analysis of their behaviour and properties. The main focus is on dynamic analysis of social networks. This paper is divided into the following parts: social networks (social network types, structural characteristics and data representations) and dynamic analysis of social networks (dynamics of social network of discussion channel, its dynamic visualisation and properties). The paper also describes the results of dynamic analysis of the data, which were collected from the Internet Relay Chat (IRC) social network. These results represent a view of moves in the framework of the social network in its entirety. The IRC network has the character of a discussion channel, which serves for solving user's mainly technical problems from the world of IT and the Unix community.

Key words

social web; web mapping; social networks; social network visualisation; dynamics visualisation; dynamic analysis of social networks

Introduction

The social web can be considered an upgrade of the classic web. The classic web can be illustrated with the idea of a worldwide billboard – anybody can publish some information piece and make it accessible for public inspection on the billboard (anybody who has the necessary skills in web page creation that is – but a considerably greater number of web users have the ability only to access published information). On the other hand, the social web or web 2.0 reinforces social interactions among users and provides an opportunity for the great majority of web users to contribute to web content. It can be said that it increases the number of web content providers. Social interactions among users are enabled by communication within social networks, the possibility to contribute to web discussions and so on. Our work focuses on social networking.

The concept of the 'social network' is frequently used and it represents one important form of the on-line social life of many people. Modern social networks help users to maintain connections and contacts to other users without dependency on a geographical locality. A social network cannot be mapped in its entirety. Mapping of a well-known social network (for example Facebook) is a difficult task when we want to visualise it in a way that guarantees readability. So, our work concentrates on small networks and also on network dynamics visualisation.

There are some known classical methods for social network analysis for off-line data processing. The data are collected within a period and data collections obtained usually do not contain a time dimension. Our work is concentrated on dynamic social network analysis. Therefore, dynamic attributes and time ordering of observations connected to the researched social network are a matter of principle. Our aim is to integrate dynamics into social networks analysis and discover some new dynamic relations.

Social networks

A social network is a set of entities (nodes, actors) that are connected. This connection can be represented by one or more types of relationship. According to Han (2003) a social network is a heterogenic and multi-relational data set that can be represented by a graph. Nodes and edges of this graph have attributes and objects (nodes) can be classified into categories. Relations can be oriented and need not be binary only. Social networks can be considered not only in a social, but also in a technological, economic or biological context. Examples of such networks can be an electrical distribution network, telecommunication networks, (computer) virus spreading, collaborative networks (co-authors), citing networks, web networks and so on. According to Repka (2011) the main concepts in social network analysis are: *actor, relation, dyad, triad, subgroup, group and network*.

An *actor* is a social entity in a network. The aim of social network analysis is to make sense of relations between these entities and to analyse the consequences of these relations. An actor can be a discrete individual, an organisation or simply a group of social entities. The actors represent the basic units of the social network in the process of its analysis.

Individual actors are connected to each other by social ties. Such a social tie can be: friendship, respect, preference, association, affiliation, interaction and so on. The collection of social ties between actors (entities) is called a *relation*. The relation $R_v \subseteq \{(A \times A) \times R\}$ between actors of the same type is a binary relation, where A is the set of actors, R is the set of relations and v_x is the intensity of the relation.

A tie between two actors is called a *dyad*. A *triad* arises by extension of the dyad by one more actor. A *subgroup* is an extension of the triad. Subgroups play an important role in the dynamic analysis of social networks. The network dynamics represent the moving of actors within the social network, that is the formation and expiration of subgroups. A group is such a set of actors as belongs to a common and finite set and which are connected with measurable ties. So *network N* can be defined as pair of the set of actors *S* and the set of all relations R over the set of actors:

$$N = (A, R); A = \{A_1, ..., A_X\}; R = \{R_1, ..., R_Y\}$$

One of the objectives of social networks analysis is identification of cohesive sub-groups and their analysis. Cohesive sub-groups are sub-sets of actors – there are strong, direct, intensive and frequent connections among them. They can be identified on the basis of the number and complexity of mutual connections, on the basis of closeness or readability of other actors of a given group and on the basis of connection frequency between actors.

Social network types

Social networks can be divided into types according various points of view. For example according to the number and type of edges (directed, undirected networks, networks with cycles and networks with multi-edges), according to the means of edge evaluation (weighted, un-weighted, marked and temporal networks), according to the number of actor types (monoecious networks contain only one type of actor, while dioecious networks contain more types of actor) and according to the number of relation types (one-relation networks, multi-relation networks). During real network modelling various network types can be distinguished: small-world phenomenon networks, scale-free networks, social cycle networks, and random and lattice networks.

Real networks are networks, which arise in a process of self-organisation. This self-organisation process can optimise the conservation of network local structure, conservation of good communication between network nodes and resistance to random errors. An example of a real network is the Internet or a network of professional contacts or telecommunication networks.

The great majority of nodes of *small-world phenomenon networks* (Kleinberg, 2000) are not connected to other nodes directly, but they do so from other nodes via a small number of edges. The small-world networks have concise shortest paths, high clustering coefficients and small node separation (these structural characteristics will be described in the next section). An interesting property of these networks is a tendency to contain cliques and n-cliques (because of the high clustering coefficient) and many high degree nodes (see next section). A great number of high degree nodes with can lead to high distribution. The *scale free networks* have distribution of a degree minimally dependent on Power Law. More details can be found in (Barabasi, 2003). *Random and lattice networks* have clustering coefficients close to zero. Its node separation grows slowly with increasing node number (Markošová, 2010). The *social cycle network* model represents networks with feedback (Douglas, 2006). Active nodes (actors) of the feedback network communicate through the network to coordinate the process of joining of new actors and to create new relations. For realisation of this process, three parameters are needed: level of actor activity, distance shortening (how soon the actor fails in the effort to find a suitable partner for new relations) and the range of searching within the network.

Structural characteristics of networks

Some properties of some social networks can be researched with the aid of their structural characteristics. The most important and useful characteristics are: node degree, shortest paths and clustering coefficient (Dorogovtsev, 2002).

The node degree k represents the total number of its connections. In the physics literature, this characteristic is usually called the ‘connectivity’, which has a different meaning in graph theory. The node degree consists of ‘in-degree’ and 'out-degree' $k = k_i + k_o$. The in-degree is the number of incoming edges of a node and the out-degree is the number of outgoing edges. It holds that:

$$P(k) = \sum_{k_i} P(k_i, k - k_i) = \sum_{k_0} P(k - k_0, k_0), \tag{1}$$

where: $P(k)$ is the degree distribution and $P(k_i, k_0$ is the joint in-degree and out-degree distribution. The in-degree distribution $P_i(k_i)$ and the out-degree distribution $P_0(k_0)$ (more obvious are notations $P(k_i)$ and $P(k_0)$ can be stated according the formula (2):

$$P_i(k_i) = \sum_{k_0} P(k_i, k_0), P_0(k_0) = \sum_{k_i} P(k_i, k_0), \tag{2}$$

If a given network has no connections to exterior space, then the average in-degree is equal to the average out-degree, see formula (3):

$$\bar{k_i} = \sum_{k_i, k_0} k_i P(k_i, k_0) = \bar{k_0} = \sum_{k_i, k_0} P(k_i, k_0), \tag{3}$$

The node degree is a local characteristic, but we shall see that degree distribution can determine some important global characteristics of random networks. The average node degree within the whole network represents the network connectivity.

The shortest path is defined as the geodetic distance of two nodes u and v, which represents the shortest one from all possible paths between these two nodes l_{uv}. The shortest path l_{uv} need not be the same as l_{vu}. The average shortest path between all node pairs within the entire network is often called the diameter of a network. It is related to the average separation of pairs of nodes.

This node separation represents another global property of a network – the closeness of nodes. The node separation l is the average shortest distance $d_{min}(a, b)$ of the nodes a and b.

$$l = \frac{1}{N} \sum_{i=1}^{N} d_{min}(a, b) \tag{4}$$

Let us consider a network with undirected edges. Then the number of all possible connections of the nearest neighbours k_i of a node u is $^{|k_i}_{2|=\frac{k_i(k_i-1)}{2}}$. Let only E_i are present. Then the clustering coefficient of the node u can be counted according to formula (4).

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{5}$$

The Average clustering coefficient over all nodes of a network is the network clustering coefficient aC.

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{6}$$

This clustering coefficient represents the probability that the two nearest neighbours of a node are also the nearest neighbours of another node.

Other parameters used in social network analysis are node numbers, edges number, nodes (edges) number in the larger weakly (strongly) connected component in the network, the number of triangles and effective average of the 90th percentile, which is the 90th percentile of the distribution of all shortest path lengths.

Network data and their representation

Social networks are sources of structural, compositional and temporal data. The structural data are obtained from connections between actors (for example friendships between people, business transactions between firms and so on). The structural data can be represented by a quadratic binary matrix (1 represents a connection and 0 represents an absence of connection). The compositional data introduce attributes of network actors, which can be direct (for example interests, age, nationality, country and so on) and indirectly obtained in the process of network analysis by counting (for example node degree, node centrality and so on). These data are static and they can be represented by an orthogonal matrix. Sometimes we are interested in the dynamics of these data, which can be researched in the process of dynamic analysis of social networks. This dynamic analysis uses temporal data, which describe the dynamic properties of a social network. The temporal data represent a time change and its time evidence.

Dynamic analysis of social networks

The dynamic analysis focuses on time changes in a researched social network. Thus the time dimension must be added into static data. Within dynamic analysis, we monitor the network development in some direction. The dynamic analysis uses visualisation methods, structural characteristics determination, machine learning methods, multi-agents modelling and combinations of these. The main problems of dynamic analysis according to Repka (2011) are:

1. searching for a suitable metric for dynamic analysis;
2. prediction of changes in social networks;
3. developing of algorithms for social network monitoring;
4. dynamic visualisation of social networks;
5. research of social networks and their characteristics in time.

Our work focuses on the last two problems. The static as well as dynamic analysis searches for suitable metrics for network description. These metrics can be used to solve the actual problem of searching for and identifying authoritative actors in a social network or in some group of the network.

In general, dynamic changes represent the creation and destruction of nodes. In real networks, the destruction of a node is often inappreciable. Such a network is called a growing network (Barabasi, 1999). Network growth can be realised both by random addition of actors and by preferential addition of actors.

The random addition of an actor is a process starting from one actor. Another actor is added to an existing actor by creating one connection to it in each time unit. Thus, the actors can be marked by the time of their coming into a network. At time t, the network has exactly t actors. So the average node degree of the node, which came into the network at time s and which we monitor at time $t+1$ is:

$$k(s,t+1) = k(s,t) + \frac{1}{t+1} \tag{7}$$

This equation can be solved and the result of its solution is formula (8):

$$k(s,t) = 1 - ln(\frac{s}{t}) \tag{8}$$

The preferential addition of actors similarly starts from one actor. In each time unit, a new actor is connected to the existing actors by creating one or more connections to existing actors in the network. The new actor selects the existing actor for connection that has the highest node degree (the highest number of existing connections). Each existing connection contributes to a node degree increase for two actors. The average node degree of the node, which came into the network at time s and which we monitor it at time t is:

$$k(s,t) = (\frac{t}{s}^{-2}). \tag{9}$$

The distribution function of node degrees is:

$$P(k, t \to \infty) = 2k^{-\gamma}, \gamma\ = 3. \tag{10}$$

The network with this distribution function has actors with a low number of connections as well as actors with a high number of connections to other actors. Such a network is called a scale-free network (Barabasi, 2003) and the model described is the Barabasi-Albert model.

Dynamics of the social network of a discussion channel

During the dynamic analysis of a network, our attention was focused on the social network IRC (Internet Relay Chat), which was created on the basis of interactions between users of a discussion channel located on the web. The advantage of this selection is that users of this discussion channel continuously communicate between each other long time, because of themes of their communication is always actual for them. These themes are related to technical problem solving, to technical news discussion or to common communication. The IRC social network is created continuously in the process of communication between users. So it is a naturally dynamic system suitable for dynamic analysis. The dynamics of this system is created by emergence. The dynamics of this discussion channel are not created purposelessly by the user but are related to the actual load of this channel according to the real need to solve some technical problems that have arisen.

From the beginning our tests concentrated on the percentage load of the researched discussion channel during particular phases. Each phase is represented by one hour of the day. Figure 1 illustrates the percentage average loading of the discussion channel in particular hours of the day.

Figure 1: Average percentage of IRC discussion channel loading. Each column represents one hour of the day. Data were collected over 211 days.

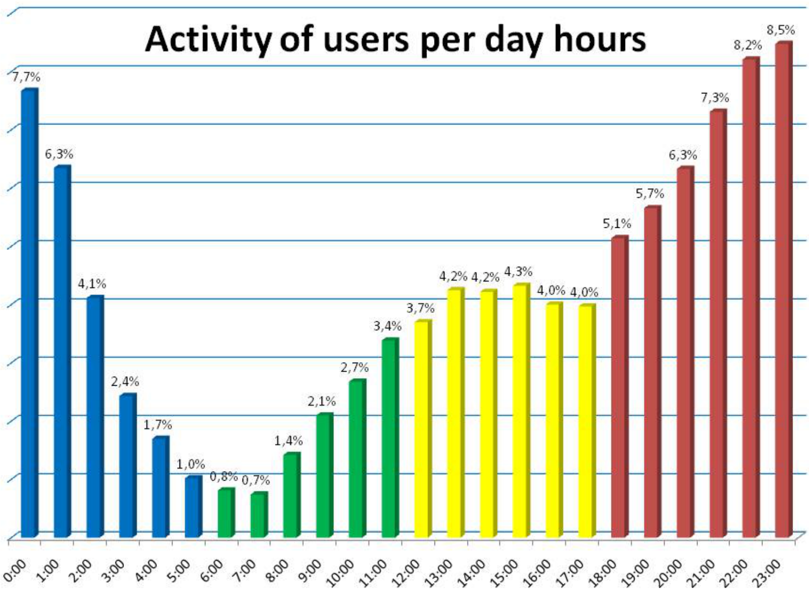


Table 1: Number of users' communications in a certain time period – each hour of the day.

Day hour Number	00:00-00:59 128,242	01:00-01:59 106,138	02:00-02:59 68,807	03:00-03:59 40,686	04:00-04:59 28,396	05:00-05:59 17,086
Day hour Number	06:00-06:59 13,563	07:00-07:59 12,391	08:00-08:59 23,755	09:00-09:59 35,083	10:00-10:59 44,776	11:00-11:59 56,598
Day hour Number	12:00-12:59 61,863	13:00-13:59 71,047	14:00-14:59 70,561	15:00-15:59 72,301	16:00-16:59 66,927	17:00-17:59 66,373
Day hour Number	18:00-18:59 85,975	19:00-19:59 94,583	20:00-20:59 105,862	21:00-21:59 122,248	22:00-22:59 137,262	23:00-23:59 141,740

Figure 2: The first iteration of the visualisation process (initialising set of social network diagram with two entities from October 09, 2009).

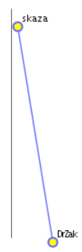
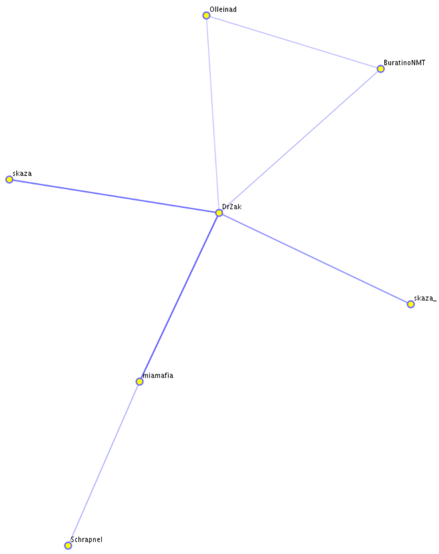


Figure 3: The 42nd iteration of the visualisation process (the social network diagram is growing – October 09, 2009).



The Figure 1 shows that the highest activity on the IRC discussion channel was at night in the evening. Table 1 gives the exact number of iterations spotted. The iterations represent communication activities between users per certain day hour. These data were collected over 211 days.

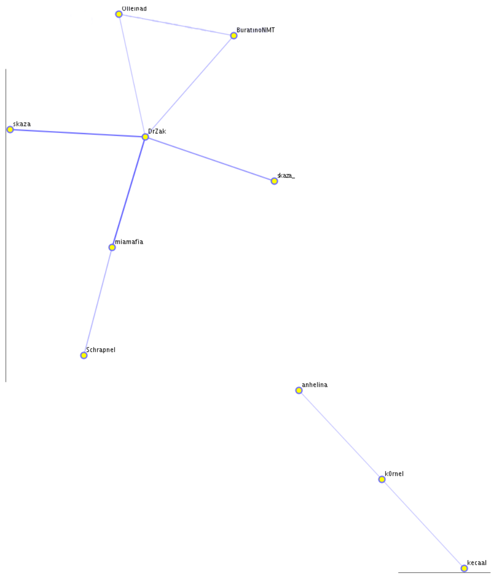
The total number of users' communications over all 211 days was 1 672 263 iterations. Activities on the discussion channel were observed with the aid of the autonomous robot Eggdrop (Eggdrop, 2011). The robot fulfils the role of a manager of users. It provides some added functionalities for operators, moderators and functionalities for safety measures.

Dynamic visualisation of social network

We have tried to visualise the IRC social network in graphical way. The graphical visualisation of the social network is a view of an actual running discussion. Each completed discussion is removed from the graph and each new discussion is added to graph by adding new nodes and edges. Running discussions are depicted by a darkening flow – line, as can be seen in the video 'czechoslovakia.wmv' available on 'http://hron.fei.tuke.sk/ rakusinec/ahReic3E/'.

This video represents running discussions and their cardinality changing over time. Such a view of a social network has a high computing complexity and with time demands on disc space increase. One possible solution to this problem is the creation of an animation that reflects only changes in time. The graph visualisation is modified only at discrete times when a change was noticed. In this way, we transformed the previously mentioned animation 'czechoslovakia.wmv' into a series of

Figure 4: The 44th iteration of the visualisation process (the social network diagram has split into two isolated parts – October 09, 2009).



8497 graph visualisations. The figures from Fig. 2 to Fig. 6 were selected from this series.

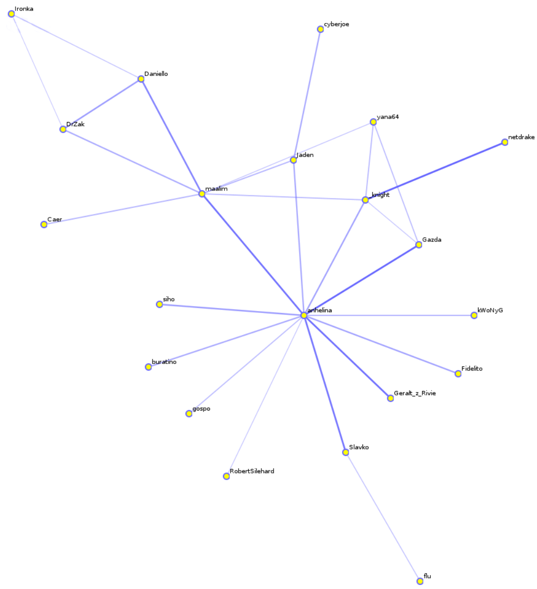
The dynamics of a social network or its evolution can be divided into two phases: an initialising phase and a monitoring phase.

The *initialising phase* starts building the social network diagram. The first interaction between the two first actors is denoted into diagram, as can be seen in Figure 2. At the beginning of building the social network diagram, the greatest changes in social network structure are denoted.

The *phase of social network monitoring* is characterised by oscillation around the stabilised graph. The oscillations depend on setting the value of a parameter that controls the time, and non-active nodes are removed from the graph after their expiration. Too low a value of this parameter enables social network monitoring nearly on-line, because the process of network development is accelerated. In this case a node is removed strictly, after very short period of inactivity. On the other hand, too high a value of this parameter causes a node to be removed after a very long period of inactivity. An extremely high value of the parameter can mean that no node is removed from graph and such a graphical visualisation can provide us with a global view of all interactions between users from the first interaction to the actual, or last one. However such a complex visualisation gradually loses its information value and readability, as illustrated in Figure 7.

The graphical visualisation of a social network and also its dynamic analysis based on graphical visualisation can be overly complex and complicated. This was our reason for focusing on the analysis of social network properties.

Figure 5: The 1778th iteration of the visualisation process (the social network diagram has become more complicated – Jun 09, 2010).



Dynamic properties of social networks

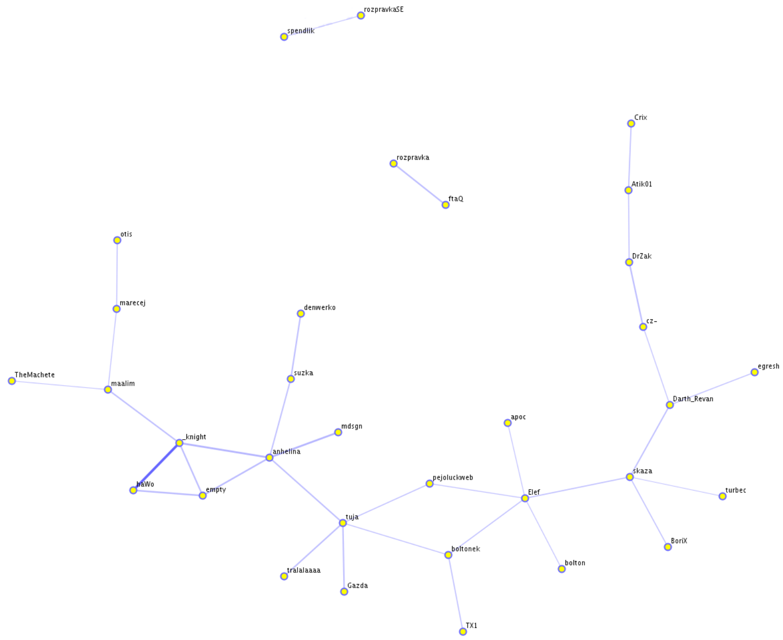
Some structural characteristics of networks were introduced within section 2.2. These static properties can be studied in their dynamics. We concentrated our attention on one of them – the node degree k, which is the most basic parameter of the investigated point (node, entity, user or actor) from the point of view of dynamic visualisation and analysis of social networks. The node degree k represents the number of all the nearest neighbours of the given node (see section 2.2).

Our experiments are related to a closed social network – the IRC discussion channel. All nodes communicate only to other nodes in the same network. The average in-degree of the node is equal to the average out-degree (see formula (3)), because the investigated network is a discussion channel, which works on the basis 'question – answer'. The set of users that did not get an answer is small and can be ignored.

We tried to research the average user-discussant and came to an average distribution of the node (user) degree depending on time. This node degree distribution is illustrated in Figure 8.

It can be seen in Figure 8 that the average node degree oscillates between values of 2 and 8. It is a relatively common value, which represents the communicative abilities of one user (actor, discussant). This value is depends also on the discussion channel load and the average number of active users-discussants during the day.

Figure 6: The 8497th iteration of the visualisation process (last social network diagram – December 05, 2010).



Conclusion

The purpose of this work was research into the dynamics of the IRC social network. We have analysed data, which were collected by PieSpy (PieSpy, 2011). The total number of collected screens was 181310. It takes up 5 GB of disc space. The social network was monitored over enough sufficient time (from October 9, 2009 to December 05, 2010) to trace permutations in the social network organisation. Clusters creation, central nodes creation and small isolated islands of small numbers of nodes were denoted.

Dynamic processes originate in this social network through emergence. This means that network users do not communicate with other users for the purpose of social network building or to gain connections, but to solve technical problems. This social network is the result of independent communication on a public level. As in any common society so in this social network the central nodes are more interesting. The central nodes – hubs – are typical, with a high number of connections to other nodes-entities. These central nodes represent very active actors, who spend more time on the discussion channel, or authoritative users, who are the best orientated in problem domains. The identification of such authoritative actors is interesting problem for us for future research.

Another promising research field is discussion analysis (Lukáč, 2008). This problem is related to discussion channels as well, but it concentrates on opinion analysis rather than dynamics analysis. This opinion analysis can be semantically

Figure 7: Overly massive visualisation of a social network (passed on from Ryze Business Networking)

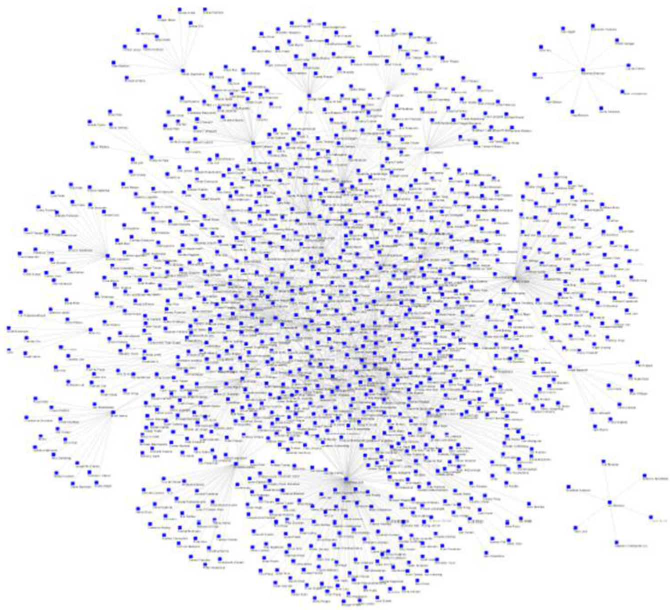
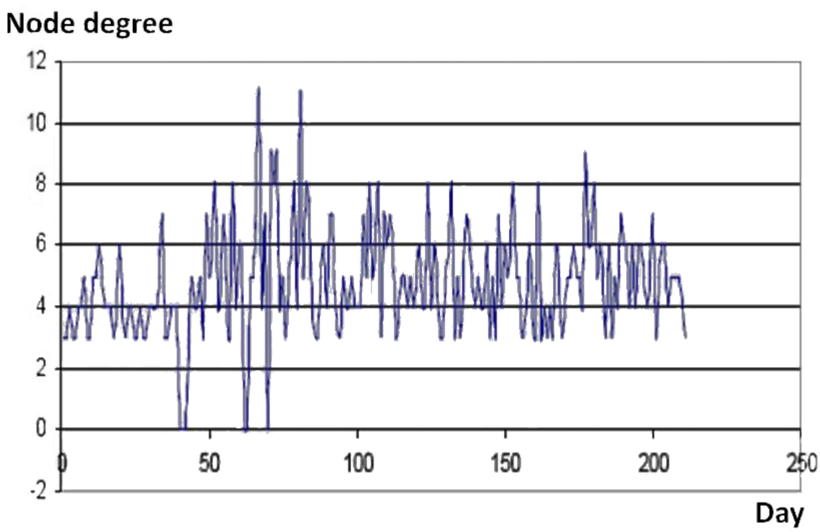


Figure 8: The distribution of the average node degree of the social network depending on time.



enhanced, which can create hybrid access to semantic and social web re-approach. This access represents social web mining from content. Social web mining from a structure can also be interesting, as introduced in Lukáč (2010) with a focus on a social network of authors and the tracking of influential concepts spreading down this network.

References

Barabasi, A. L. (2003): Scale-free Network. *Scientific American*, 288, pp. 60–69.

Barabasi, A. L., Albert, R. (1999): Emergence of scaling in random networks. *Science*, 286, pp. 509–512.

Dorogovtsev, S. N., Mendes, J. F. F. (2002): Evolution of Networks. *Adv. Phys*, Vol. 51, pp. 1079–1187.

Douglas, R. a kol. (2006): Generative Model for Feedback Networks. In: *Physical Review E.* , pp. 016119-1–016119-8.

EGGDROP. (2011): The world's most popular open source Internet Relay chat bot [on-line]. [cit. 2011-04-11]. Available at: `http://www.omgirc.com/index.php?itemid=26`.

Hann, W. P. (2003): Hyperlink Network Analysis: A New Method for the Study of Social Structure on the Web Royal Netherlands Academy of Arts & Sciences, Amsterdam, The Netherlands. *Connections*, Vol. 25, No. 1, pp. 49–61.

Kleinberg, J. (2000): The Small-world Phenomenon: An Algorithmic Perspective. In: *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing.* New York : The Association for Computing Machinery, pp. 163–170. (ISBN 1-58113-184-4.)

Lukáč, G. (2010): A Proposal for an Approach to Extracting Conceptual Descriptions of Hyper-linked Text Documents. *Int. J. on Social Media MMM: Monitoring, Measurement and Mining*, Vol. 1, No. 1, pp. 98 to 101. (ISSN 1804-5251.)

Lukáč, G., Butka, P., Mach, M. (2008): Semantically-enhanced extension of the discussion analysis algorithm in SAKE. In: *Proceedings of the SAMI 2008.* Budapest : Budapest Tech, pp. 241–246.

Markošová, M. (2010): Networks Dynamics. In: Kvasnička a kol. (ed.) *Artificial Intelligence and Cognitive Sciences II.* Bratislava (SK) : STU, pp. 321–377.

PieSpy. (2011): PieSpy Social Network Bot. [on-line]. [cit. 2011-04-14]. Available at: `http://www.jibble.org/piespy/`.

Repka, M. (2011): *Analysis of selected types of social networks.* Košice (SK) : Technical University in Košice, 75 pp.

RBN. (2011): Ryze Business Networking [on-line]. [cit. 2011-04-14]. Available at: `http://www.ryze.com/index.php`.

This work was supported by the Slovak Grant Agency of the Ministry of Education and Academy of Sciences of the Slovak Republic within project 1/0042/10 'Methods for identification, annotation, search, access and composition of services using semantic metadata in support of selected process types' (50%). This work is also the result of the implementation of the project Development of the Centre of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (50%).