Deliverable 5.2.5
Degeneracy and redundancy in self-organised systems

Due date of deliverable:	December 2007
Actual submission date:	December 2007
Dissemination level:	PU – public
Work Package 5.2:	Evolved Tinkering and Degeneracy as Engineering Concepts
Participants:	UPF UniBO Telenor
Authors of deliverable:	Sergi Valverde (svalverde@imim.es) Ricard V. Solé (ricard.sole@upf.edu)

Abstract

This report comprises the complete D5.2.1 deliverable as specified for workpackage WP5.2 in Subproject SP5 of the DELIS (Dynamically Evolving Large-scale Information Systems) Integrated Project.

The essential goal of the DELIS project is to understand, predict, engineer and control large evolving information systems. The main aim of this workpackage is to understand how evolved structures emerge in networks when there is no central design or control.

Complex networks emerge under different conditions including design (i.e., top-down decisions) through simple rules of growth and evolution. Such rules are typically local when dealing with biological systems and most social webs. An important deviation from such scenario is provided by groups, collectives of agents engaged in technology development, such as open source (OS) communities. Here we analyze their network structure, showing that it defines a complex weighted network with scaling laws at different levels, as measured by looking at e-mail exchanges. We also present a simple model of network growth involving non-local rules based on betweenness centrality. Our weighted network analysis suggests that a well-defined interplay between the overall goals of the community and the underlying hierarchical organization play a key role in shaping its dynamics.

Contents

1 Introduction

Networks predate complexity, from biology and society to technology [1]. In many cases, large-scale, system-level properties emerge in a self-organized manner from local (bottom-up) interactions among network components. This is consistent with the general lack of global goals that pervade cellular webs or acquaintance networks. However, when dealing with human collective efforts towards a given objective, such as in a company or in distributed technology development, the situation can be rather different. Top-down decisions might dominate the structure and function in a hierarchical way. But how to distinguish between the two scenarios?

The intrinsic network organization of social interactions allows to explore this questions in depth. Many of these networks can be reconstructed by using e-mail exchanges among agents . The resulting graph provides a well-defined picture of the global community organization. By looking at its topology, we could in principle identify the presence (or absence) of self-organized (SO) or designed (top-down) patterns. Here SO refers to patterns emerging from local rules. Such system would display global features resulting from a bottom-up dynamics. Eventually, a model of network growth can be proposed in order to explain the origin of such pattern. An example of this is the work by Caldarelli et al. [2] who studied the emergence of weighted social networks. These authors showed that the structure of e-mail webs could be explained using a simple local mechanism based on positive feedback and reciprocity.

In this paper we explore the problem of how SO and hierarchy might actually emerge and coexist in a distributed community of technological developers. Specifically, we will present the first analysis of weighted open source (OS) communities [3]. In OS communities, software is developed through distributed cooperation among many agents. These communities are known to display a large amount of distributed, bottom-up organization. Specifically, large groups of programmers are involved in building, assembling and specially maintaining large-scale software structures. The community plays multiple roles as a design system but also as a distributed intelligence system able to accept or reject changes introduced by agents. As described, it looks like we are talking about a largely self-organized entity. Given the quality of the information available on their internal structure, OS organizations offer a unique opportunity to test if they are fully self-organized social groups [4] in constrast with more hierarchical, top-down organized social groups (i.e., large companies).

One possible test to these potential modes of community organization involves using the network of interaction between programmers working in a given software system. Software systems are themselves complex networks [5], which have been shown to display small world and scale-free architecture. Since the topological organization of software designs is scale-free, we might suspect that the community organization also displays common traits with the underlying software architecture. Previous work on engineering problem-solving networks involved in product development [6] revealed that these groups define a complex network with heterogeneous link distributions. However, these networks are unweighted and largely dominated by top-down constraints. Here, we consider a different type of engineering community where relations among agents are weighted and change in time without previously defined hierarchies.

As we will shown here, OS networks (OSN) display scaling laws but also a well-defined core of main programmers defining a special subset of agents. Such finding suggests that, even in these distributed groups of individuals, emergence of hierarchy might be inevitable. Our analysis reveals the interplay between bottom-up, distributed decision making periphery in the OSN involving many agents and a top-down driven, centralized core of agents. Such rich-club structure seems to place some limits to the degree of distributedness achievable by multiagent-based technological design.

In the next sections we summarise, in overview, results given in [7]. There we provided an empirical analysis of OS developer networks and presented a network growth model that agrees with the empirical data.

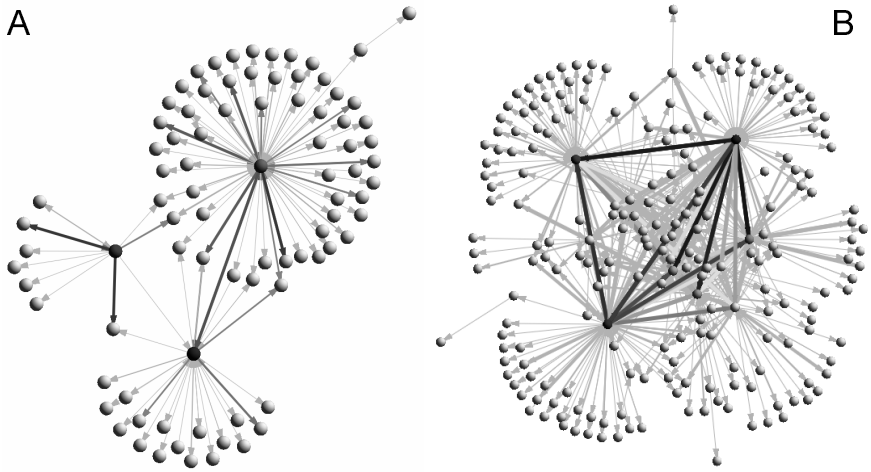

Figure 1: Social networks of e-mail exchanges in open source communities. Line thickness represents the number of e-mails flowing from the sender to the receiver. Dark depicts active members and frequent communication. (A) Social network G_{Amavis} for the Amavis open-source community. (B) Social network G_{TCL} for the TCL (i.e., Tool Command Language) open-source community with $N = 215$ members and $\langle k \rangle \approx 3$. In both networks, a few hubs (center dark nodes) route the bulk of information generated by many periphery nodes.

2 Empirical data from e-mail networks of open source communities

We have analyzed 120 OS networks corresponding to different software projects. We reconstruct the social network with the following method. For each OS network $\Omega = (V, L)$, nodes $v_i \in V$ depict community members while directed links $(i, j) \in L$ denote e-mail communication whether the member i replies to the member j. At time t, a member v_i discovers a new software error (bug) and sends a notification e-mail. Afterwards, other members investigate the origin of the software bug and eventually reply to the message, either explaining the solution or asking for more information. Here $E_{ij}(t) = 1$ if developer i replies to developer j at time t and is zero otherwise. From E_{ij} we define link weight e_{ij} as the total amount of e-mail traffic flowing from developer i to developer j:

$$e_{ij} = \sum_{t=0}^{T} E_{ij}(t) \tag{1}$$

where T is the timespan of software development. We have found that e-mail traffic is highly symmetric, i. e. $e_{ij} \approx e_{ji}$. Thus, we can make the simplifying assumption that the network is undirected.

However, we do not restrict our study to purely topological links. Instead, their weighted structure is also taken into account. The edge weight (interaction strength) is defined as $w_{ij} = e_{ij} + e_{ji}$, which provides a measure of traffic exchanges between any pair of members. From this weighted matrix we can estimate node strength [10] as a local measure defined as:

$$s_i = \sum\nolimits_j w_{ij} \tag{2}$$

i. e. the total number of messages exchanged between node i and the rest of the community. This definition will be used below in our analysis of the weighted OS network.

Figure 1 shows two social networks recovered with the above method. We can appreciate an heterogeneous pattern of e-mail interaction, where a few members handle the largest fraction of e-mail

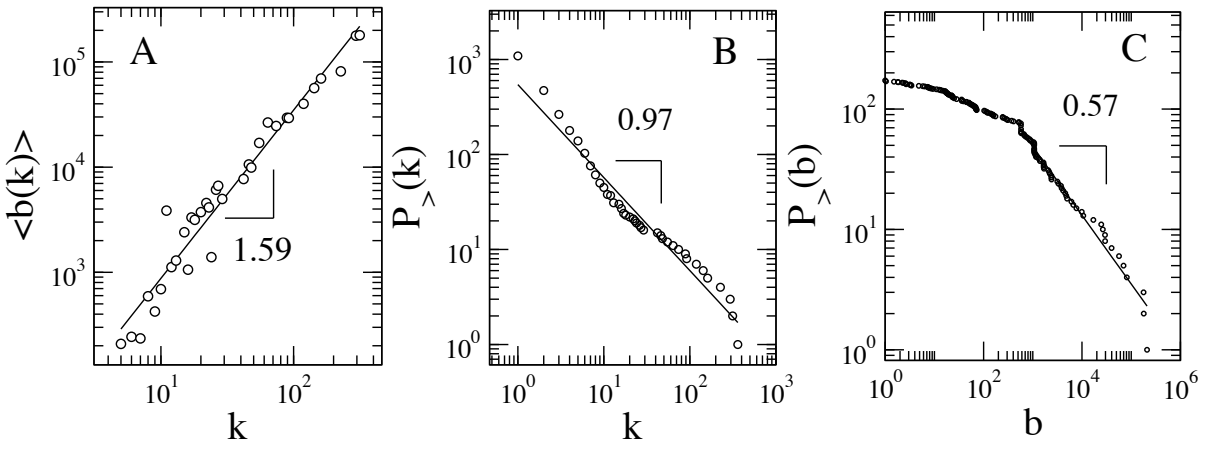

Figure 2: (A) Average betweeness centrality scales with degree $\langle b(k)\rangle \sim k^{\eta}$ with $\eta \approx 1.59$ for the Python OS community. This exponent is close to the theoretical prediction $\eta_{BA} \approx (\gamma - 1)/(\delta - 1) = 1.70$. (B) Cumulative distribution of undirected degree $P_{>}(k) \sim k^{-\gamma+1}$ with $\gamma \approx 1.97$. (C) Cumulative distribution of betweeness centrality $P_{>}(b) \sim b^{-\delta+1}$ with $\delta \approx 1.57$ for $b > 10^2$.

traffic generated by the OS community. The undirected degree distribution is roughly a power-law $P(k) \sim k^{-\gamma}$ with $\gamma \approx 2$ (see fig. 2B). However, $P(k)$ displays a hump at some intermmediate degree k_c (see fig. 2B). The hump suggests a two-level classification of nodes in the OS network: periphery nodes with few connections having $k < k_c$ and hub nodes having $k > k_c$. This desviation might be an indication of a rich-club ordering in the OS network (see below).

In order to understand the role played by hubs in OS networks, we have measured the betweeness centrality b_i (or node load [8]), i.e. the number of shortest paths passing through the i-th node [9]. Betweeness centrality displays a long tail $P(b) \sim b^{-\delta}$ with an exponent δ between 1.3 and 1.8 (see table I and also fig. 2C). It was shown that betweeness centrality scales with degree in the Internet autonomous systems and in the Barabási-Albert network [11], as $b(k) \sim k^{-\eta}$. From the cumulative degree distribution, i. e.

$$P_{>}(k) = \int_{k}^{\infty} P(k)dk \sim k^{1-\gamma} \tag{3}$$

and the corresponding integrated betweenness, with $P_{>}(b) \sim b^{1-\delta}$, it follows that $\eta = (\gamma-1)/(\delta-1)$ [12]. The social networks studied here display a similar scaling law with an exponent η slightly departing from the theoretical prediction (see fig. 2A and table I). The strong correlation between node load and large degree indicates that hubs tend to dominate e-mail discussions in the OS community.

3 Rich-clubs

In order to discover if some programmers are more significant than others we define a weighted rich-club coefficient $\Phi(S_k, k)$ as follows:

$$\Phi(S_k, k) = \frac{W_S(k)}{E_S(k)\langle w\rangle} \tag{4}$$

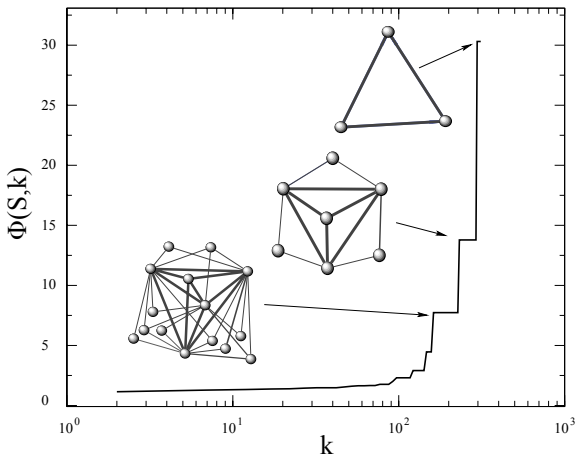

Figure 3: Plot of the weighted rich-club coefficient $\Phi(S,k)$ against node degree k for the Python OS network. There is a significant deviation for $k > 10^2$ that signals the rich-club ordering for this particular community. The subgraphs show the k-scaffolds or the predicted rich-clubs for different degrees $k > 100$. Line thickness indicate the weight attached to the link. We can appreciate how three nodes have a much more stronger internal interaction (i.e., exchange a larger number of e-mails) than with the rest of nodes.

where $E_S(k)$ depicts the number of edges in the k-scaffold of the OS network, $\langle w \rangle = 1/E \sum_{ij} w_{ij}$ is the average edge weight for the full network, E is the total number of edges, and $W_S(k) = \sum_{i,j \in S(k)} w_{ij}$ is the sum of edge weights linking nodes in the k-scaffold subgraph [13]. The coefficient signals any deviation from an homogenous distribution of weights in the k-scaffold. When weights are distributed at random then both the numerator and denominator will be the same and $\Phi(S,k) \approx 1$. However, it is easy to see that inhomogeneities in the weight distribution among edges (i.e., when large weights are clustered in the edges of some connected subgraph) yield $\Phi(S,k) \gg 1$. This seems to be the case for OS networks (see fig. 3) where a dramatic growth of $\Phi(S_k,k)$ is observed when the core set of programmers is reached. Such divergence clearly reveals the non-homogeneous nature of the OSN, where a large fraction of e-mails flow through a few OS hubs.

4 Predictive social network simulation model

We present a top-down model that predicts the evolution and dynamics of the OS network, including the (undirected) degree distribution $P(k)$ and measurements of local correlations (see fig.4C, fig.4D, and fig.4E). This model is motivated by three empirical observations: (i) there is a non-lineal relationship between node strength and degree (previouly reported in [18]). In a related paper, this relationship has been explained with a betweenness centrality model [14]. (ii) Betweenness centrality strongly correlates with node strength (see fig. 4A). (iii) OS networks have a rich-club core (see above). The rich-club indicates a characteristic scale in the system that emerges from an external reinforcement of core members' activities.

Core members will be more frequently e-mailed because of their importance. Key agents keep the community as a coherent system. In this context, agents exploit social cues to evaluate one another's social status [15]. A natural surrogate of social status is the number of e-mails posted (and received) by the member, i.e., node strength s_i (see section II). Members earning high social status are arguably the most visible and thus, they will be accessed much more frequently [16]. These key members have a global picture of the whole system, instead of being aware of just some specific parts

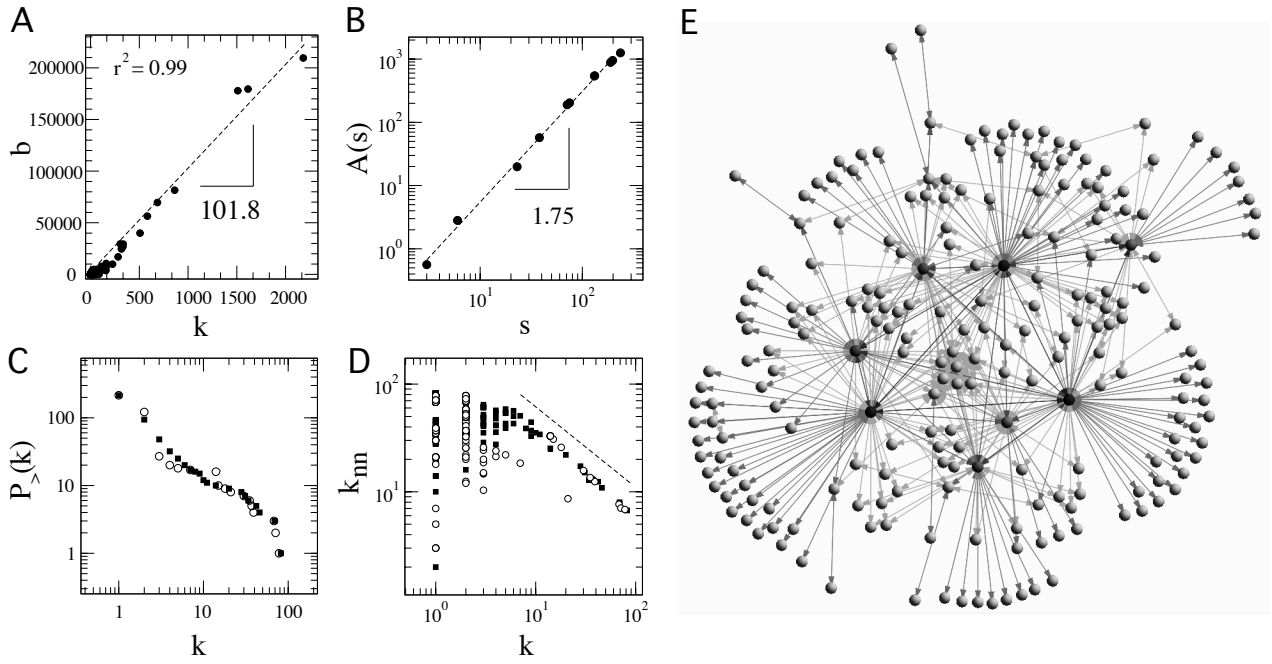

Figure 4: Social network simulation (A) Linear correlation between node strength s_i and betweeness centrality (or node load) b_i in the Python community. The correlation coefficient is 0.99. This trend has been observed in all communities studied here. (B) Estimation of α in the TCL community (see text). (C) Cumulative degree distribution in the simulated network (open circles) and in the real community (closed squares). All parameters estimated from real data: $N = 215$, $m_0 = 15$, $\langle m \rangle = 3$ and $\alpha = 0.75$. (D) Scaling of average neighbors degree vs degree in the simulated network (open circles) and in the real social network (closed squares). There is very good overlap between model and data for large k. (E) Rendering of the simulated OS network Ω to be compared with the OS network G_{TCL} in fig. 1B.

of it. Members having a deeper knowledge of the overall system's architecture are likely to process high amounts of information. If we think in terms of agents in a network, we should expect them to canalize information flowing from many different parts of the network [17].

Taking into account the above, the algorithm for evolving the (undirected) social network $\Omega = (V, L)$ consists of the following stages: (i) The system starts (as in real OS systems) from a small fully-connected network of m_0 members. (ii) A new member j joins the social network at each time step. The new member reports a small number of an average $\langle m \rangle$ new e-mails (iii) For each new e-mail, we determine the target node by a non-local preferential attachment rule. The probability that new member j sends an e-mail to an existing member i is proportional to node betweenness b_i, or alternatively, to the node strength s_i

The networks generated with the previous model are in very good agreement to real OS networks. For example, fig. 4 compares our model with the social network of TCL software community.

5 Conclusion

Our analysis shows that open source communities are closer to the Internet and communication networks than to other social networks (e.g., the network of scientific collaborations). The social

networks analyzed here are dissasortative from the topological point of view and assortative when edge weights are taken into account. This is consistent with the absence of topological rich-club that is nonetheless detected when link weights are taken into account. The rich-club phenomenon in OS networks seems to be related to a pattern of non-local evolution. Such a non-local component appears to be related with the presence of a core of programmers that make decisions based on a global view of the system. Core programmers would both introduce a top-down control and receive a large amount of e-mail traffic from secondary members. Based on these ideas, we have presented a model that predicts many global and local social network measurements of the OS network.

References

[1] Dorogovtsev, S. N. & Mendes, J. F. F., *Evolution of Networks: From Biological Nets to the Internet and WWW*, Oxford University Press, New York (2003); Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U., *Physics Reports*, 424 (2006) 175; Newman, M. E. J., *SIAM Review* 45, (2003), 167-256; Albert, R., and Barabási, A.-L., *Rev. of Mod. Phys.* 74 , (2002) 47.

[2] Caldarelli, G., Coccetti, F., and de Los Rios, P., *Phys. Rev. E.*, **70**, 027102 (2004).

[3] Raymond, E. S., *First Monday*, **3** (1998).

[4] Ball, P., *Critical Mass: How one thing leads to another*, Arrow Books, (2004).

[5] Valverde, S., Ferrer-Cancho, R. & Solé, R. V. *Europhys. Lett.* **60** (2002) 512-517; Myers, C. R., *Phys. Rev. E* 68, 046116 (2003); Valverde, S. and Solé, R. V., *Phys. Rev. E* , **72**, 026107 (2005); Valverde, S., and Solé, R. V., *Europhys. Lett.*, **72**, 5, pp. 858–864 (2005).

[6] Braha, D., and Bar-Yam, Y., *P*hys. Rev. E., **69**, 016113, (2004).

[7] Self-Organization versus Hierarchy in Open Source Social Networks Sergi Valverde and Ricard V. Sol Physical Review E 76, 32767 (2007)

[8] Goh, K.-I., Kahng, B., and Kim, D., *Phys. Rev. Lett.*, **87**, 278701 (2001).

[9] Brandes, U., *Journal of Mathematical Sociology*, **25**, 2, pp. 163–177 (2001).

[10] Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A., *Proc. Natl. Acad. Sci. USA*, **101**, 3747 (2004).

[11] Barabási, A.-L., and Albert, R., *Science*, **286**, 509 (1999).

[12] Vazquez, A., Pastor-Satorras, R., and Vespignani, A., *Phys. Rev. Lett.*, **65**, 066130 (2002).

[13] Here, S_k denotes the naked minimal k-scaffold subgraph.

[14] Goh, K.-I., Kahng, B., and Kim, D., *Phys. Rev. E.*, **72**, 017103 (2005).

[15] Stewart, D., *American Sociological Review*, **70**, pp. 823-842 (2005).

[16] Watts, D. J., Sheridan Dodds, P., and Newman, M. E. J., *Science*, **296**, 5571, (2002) 1302 - 1305.

[17] Sheridan Dodds, P., Watts, D. J., and Sabel, C. F., *PNAS* **100** (21), (2003) 12516-1252.

[18] Sergi Valverde, Guy Theraulaz, Jacques Gautrais, Vincent Fourcassié, and Ricard V. Solé, "Self-Organization Patterns in Wasp and Open-Source Communities", IEEE Intelligent Systems, vol. 21, no. 2, pp. 36-40, Mar/Apr, 2006. [DELIS-TR-0434]

[19] Lada A. Adamic, Orkut Buyukkokten, and Eytan Adar, "A social network caught in the Web", First Monday, 8(6), June (2003).

[20] A.-L. Barabási, "The origin of bursts and heavy tails in human dynamics", Nature 435, 207211 (2005).

[21] Kevin Crowston and James Howison, "The Social Structure of Free and Open Source Software Development", First Monday, February (2005).

[22] Jin Xu, Yongqin Gao, Scott Christley, and Gregory Madey, "A Topological Analysis of the Open Source Software Development Community", Proc. IEEE 38th Hawaii Int. Conf. Syst. Sci, (2005).

[23] M. E. Conway, "How Committees Invent?", Datamation, 14 (4), pp. 28-31, (1968).

[24] M. Anghel, Zoltán Toroczkai, Kevin E. Bassler, and G. Korniss, "Competition-Driven Network Dynamics: Emergence of a Scale-Free Leadership Structure and Collective Efficiency", Phys. Rev. Lett. 92, 058701 (2004)

[25] Sergi Valverde and Ricard V. Solé, "Self-organization and Hierarchy in Open-Source Social Networks", Submitted to Physical Review E (2006). Previous preprint: http://arxiv.org/abs/physics/0602005. [DELIS-TR-0433]

[26] Duncan J. Watts, Peter Sheridan Dodds, and M. E. J. Newman, "Identity and Search in Social Networks", Science 296 (5571), 1302 (2002).

[27] Peter Sheridan Dodds, Duncan J. Watts , and Charles F. Sabel, "Information exchange and the robustness of organizational networks", PNAS, 100(21), pp. 12516-12521, (2001).

[28] Lerner, J., and Tirole, J., *Journal of Industrial Economics*, **52**, pp. 197-234, (2002).

[29] Zhou, S. and Mondragon, R.J., IEEE Comm. Lett., 8, pp. 180182, (2004).

[30] Mockus, A., Fielding, A., Herbsled, J., in Proc. Int. Conf. Soft. Eng., Limerick, Ireland, pp. 263-272, (2002).

[31] S. Wasserman, and K. Faust, "Social network Analysis: Methods and Applications", Cambridge University Press, Cambridge (1994)