Social Network Analysis of Developers' and Users' Mailing Lists of Some Free Open Source Software

Armel Jacques Nzekon Nzeko'O, Matthieu Latapy, Maurice Tchuente

Social network analysis of developers and users mailing lists of some free open source software

Armel Jacques Nzekon Nzeko'o[1,2]
armeljanz@gmail.com

Matthieu Latapy[3]
matthieu.latapy@lip6.fr

Maurice Tchuente[1,2]
maurice.tchuente@gmail.com

(1) IRD UMI 209 UNMISCO, Laboratoire d'Informatique et Applications, Université de Yaoundé I, BP 337 Yaoundé, Cameroun
(2) LIRIMA, Equipe IDASCO, Faculté des Sciences, Département d'Informatique, BP 812 Yaoundé, Cameroun
(3) Sorbonne Universités, UPMC Université Paris 06, UMR 7606, LIP6, F-75005 CNRS, UMR 7606, LIP6, F-75005, Paris, France

Abstract— **As reported by Kevin Crowston and co-authors in a recent paper, free open source software is a very important social phenomenon that involves nearly one million programmers, a myriad of software development firms, millions of users, and its financial impact is huge since for instance the cost of recreating available free software is estimated in tens of billions of euros. Free open source software projects generally have one mailing list for developers and another one for users. This large number of mailing lists changes constantly and shows a great variety with respect to membership and topics covered. This makes them very difficult to monitor. One way of overcoming this Big Data Challenge is to identify some easily computable global indicators that can be used for instance to detect important events. We illustrate this approach here by making a social network analysis and comparison of developers' and users' mailing lists of four free open source software projects: CentOS, GnuPG, Mailman and Samba. We show that these mailing lists have some common characteristics : the number of messages, the time durations and the interlink times can be fitted using power and lognormal laws with suitable scales and parameters ; for the interlink time, the analysis is done using the temporal delta density inspired by the delta density introduced by Viard and Latapy. This similarity between the characteristics of mailing lists also applies to the structure of dominant groups. For the time evolution of the number of messages, GnuPG exhibits a particular behavior. The interpretation of the different parameters gives very interesting insights into the membership and the type of topics covered by the mailing lists. The analysis carried out here and similar studies cited in this paper can therefore be considered as a first step towards the designing of building blocks for monitoring mailing lists.**

Keywords— Big data; free open source software project; mailing list; discussion thread; dominant member; power law; lognormal law; complementary cumulative distribution; delta density; characteristic time; outstanding event detection.

I. INTRODUCTION

An electronic mailing list is a particular usage of email that allows members of a group to communicate easily. When a member of the list sends a message, he/she uses the group's special address and the e-mail is broadcast to all the members of the mailing list.

Free open source software projects generally have one mailing list for developers and another one for users. These two mailing lists, considered globally, have been studied for the structural characteristics of the underlying networks [1, 2, 3], discussions themes [4], member's activities [5, 6, 7] and outstanding events [8, 9]. Other studies have separately analyzed developers' and users' mailing lists with respect to key-participants [5, 9], knowledge sharing, posting and replying activities [10, 11, 12].

In this paper, we make a social network analysis and comparison of developers' and users' mailing lists of four free open source software projects: the operating system CentOS, the free implementation of OpenPGP, GnuPG, that allows to encrypt and sign data and communication, the GNU mailing list manager Mailman and Samba, a software that allows machines under Unix system to manage printers and files. More precisely, we are interested in the distribution of messages in mailing list threads, the distribution of threads durations, the relationship between duration and number of messages, the proportion of common dominant members, the temporal characteristics of mailing list threads and the detection of outstanding events.

The rest of this article is organized as follows. Section 2 gives some basic notations and definitions. Section 3 presents the data used as well as some preprocessing applied to these data. Section 4 is devoted to empirical results. We present some related work in section 5 and the conclusion in section 6.

II. BASIC NOTATIONS AND DEFINITIONS

A. Tools for the description of empirical distributions of thread properties

In this subsection, we recall some social network tools commonly used to study empirical distributions of thread properties. More precisely we are interested in classical fit methods and goodness of fit evaluations. The two functional forms used here to explain the observed cumulative distributions of empirical data are the power law (PL) that is very common in social network analysis and the lognormal law (LN) hypothesis. After selecting the optimal values of the parameters for both hypotheses, we compare the

978-1-4673-7278-7/15 $31.00 © 2015 IEEE
DOI 10.1109/BigDataCongress.2015.119

relevance of the fits by comparing the empirical distribution to each fit directly. A classical way to do it is the Kolmogorov-Smirnov (KS) distance.

Many integer features in social networks follow a power law i.e. the probability that the feature has a value k is $p_k \sim k^{-\alpha}$ for some constant α generally between 2 and 3. In an early paper, Price [13] was the first to observe power law distributions in social networks, as he was studying the in-degrees and the out degrees of citation networks.

An alternate way of studying such integer features is to use the complementary cumulative distribution function (CCDF) $P_k = \sum_{k'=k}^{\infty} p_{k'}$ which represents the probability that an entity has a feature value greater than or equal to k. A very interesting property of a power law distribution is that $P_k \sim k^{-(\alpha-1)}$ [14]. As noted by Barabasi and Albert [15], power law property may be a consequence of a robust self-organizing mechanism: networks expand continuously by the addition of new entities and, new entities attach preferentially to sites that are already well connected, i.e. rich nodes get richer.

Clauset et al. [16] have noted that in practice, few empirical phenomena obey power laws for all values of x. More precisely, the power law often applies only for values greater than some minimum x_{min}.

A positive variable X is said to have a lognormal distribution if the random variable $Y = \log(X)$ has a normal distribution. The density function for a lognormal distribution is $LN(x;\mu,\sigma) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$.

B. Temporal Δ-density of link streams

Following [17], we model a discussion thread as a link stream $L = (l_i)_{i=1..k}$ with $l_i = (t_i, u_i, v_i)$, that contains k messages. We call $T = t_k - t_1$ its duration. The i-th inter-link time τ_i is defined by $\tau_i = t_{i+1} - t_i$, for $i = 1,\ldots, k-1$.

Suppose a Δ between 0 and T is given. We define the instantaneous Δ-density of L at time t as

$$\delta_\Delta^t(L) = \begin{cases} 1 \; if \; \exists t_i \in [t-\Delta, t] \\ 0 \; otherwise \end{cases}$$

This instantaneous Δ-density tells us about the occurrence of a message in the time-interval $[t-\Delta, t]$, and is defined for $t \geq \Delta$. Let us denote by $\delta_\Delta(L)$ the probability that a randomly chosen time-interval of size Δ contains at least one message. It can be shown that

$$\delta_\Delta(L) = \frac{\int_{\alpha+\Delta}^{\omega} \delta_\Delta^t(L).\,\mathrm{dt}}{\mathrm{T}-\Delta}$$

where $\mathrm{T}-\Delta$ is a normalizing coefficient. The discrete form of temporal Δ-density is [17]

$$\delta_\Delta(L) = 1 - \frac{\sum_{i=1}^{k-1} \max(\tau_i - \Delta, 0)}{T-\Delta}.$$

C. Characteristic inter-link time of a discussion thread

In [17], Viard et al. suggested a procedure to identify relevant values of Δ that may reveal the dynamics of links in a sub-stream. In this paper we use the same procedure as follows: first compute the temporal Δ-density for various values $\Delta_i = \beta^i$, with $\beta > 1$, where $i = 1, 2\ldots$ and Δ_i remains between 1 second and the whole duration of the thread; then observe the growth of $\delta_\Delta(L)$ as a function of Δ. The value of Δ which causes the greatest growth is called the characteristic inter-link time of the corresponding stream, and is denoted $\tau(L)$.

III. DATASET

The archives that we used are all from mailing lists maintained by the Mailman mailing list manager. In this system, communications are archived by month in a text file with Internet Message Format (RFC 5322). For each software project, we have downloaded the archive corresponding to the mailing lists for developers and for users, for the period from January 2010 to December 2014.

After gathering the 60 files of each of the 8 archives mailing lists, we merged them into a single file in chronological order. In the resulting file, each message m has one author $a(m)$ and a publication date $t(m)$. The e-mail may be a response to a previous message $p(m)$. Some messages are not answers to any previous message; in this case, $p(m) = m$. Such messages are called root messages.

Each root message m induces a thread which corresponds to a set $T(m)$ of messages such that: m belongs to $T(m)$, all answers to m belong to $T(m)$ and any answer to an answer to m belongs to $T(m)$. To have a representative subset of threads, we deleted all messages that were answering to messages absent from the archive.

Table I below provides details on the data used for each mailing list studied here. We have the project name in the first column, the type of mailing list in the second column. The next columns contain the number of threads, the number of participants and the number of messages.

TABLE I. STATISTICS ON DATA.

	Type	***Threads***	***Participants***	***Messages***
CentOS	*devel*	875	654	5 967
	users	9 080	2 717	47 735
GnuPG	*devel*	940	314	2 539
	users	2 364	1 396	9 949
Mailman	*devel*	518	249	2 235
	users	2 272	1 299	7 050
Samba	*devel*	7 144	1 711	22 953
	users	9 865	4 716	20 543

IV. EMPIRICAL RESULTS

A. Distribution of messages in mailing list threads

For the three projects CentOS, Mailman and Samba, the proportion of threads with a great number of messages is more important for developers than for users. This may be due to some hot topics leading to intense debates among developers whereas in users' communities the questions raised are usually answered by experts after few messages. For GnuPG we have the opposite situation, meaning that

intense debates are more frequent within users than in the group of developers.

These observations follow from table II because in the power law fitting, when α is smaller, the probability of having threads with large number of messages is larger.

In columns 6 and 9 of table II the low value of Kolmogorov-Smirnov test (KS) show that the fitting with power law for values greater than x_{min} and with lognormal law for the entire CCDF are quite accurate. This accuracy is illustrated in figure 1 for CentOS.

TABLE II. POWER LAW AND LOGNORMAL LAW FIT OF THE CCDF OF THE NUMBER OF MESSAGES.

		Power law				*Lognormal law*		
Project	***Type***	x_{min}	**α**	***disc.***	***KS***	**μ**	**σ**	***KS***
CentOS	*devel*	13	2.6	81%	0.05	1.2	1.3	0.06
	users	20	1.2	94%	0.02	1.3	1.1	0.02
GnuPG	*devel*	7	2.5	74%	0.07	0.8	1.3	0.06
	users	7	3.1	88%	0.02	0.4	1.1	0.05
Mailman	*devel*	10	2.8	88%	0.03	0.7	1.2	0.02
	users	15	4.0	96%	0.02	0.5	1.1	0.02
Samba	*devel*	7	3.0	83%	0.05	0.8	1.0	0.04
	users	8	2.5	80%	0.03	0.8	1.3	0.03

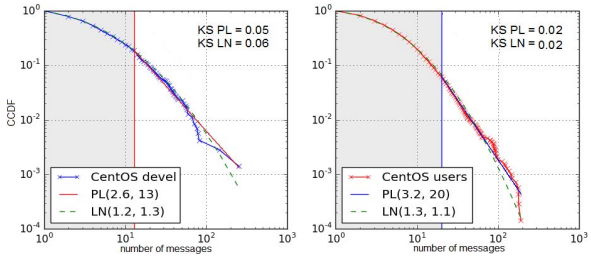

Figure 1. Complementary cumulative distribution of the number of messages fitted by power law and lognormal law in log-log scale.

B. Cumulative distribution of threads durations

The cumulative distributions of threads durations in mailing lists follow lognormal laws with the parameters and accuracies shown in table III and illustrated in figure 2 for GnuPG.

TABLE III. LOGNORMAL LAW FIT OF THE CUMULATIVE DISTRIBUTION OF THREAD DURATIONS.

	CentOS		GnuPG		Mailman		Samba	
	dev	*usr*	*dev*	*usr*	*dev*	*usr*	*dev*	*usr*
μ	*2.9*	2.7	3.3	3.0	3.2	2.8	2.9	2.9
σ	*0.4*	0.4	0.3	0.3	0.3	0.3	0.4	0.4
KS	*0.12*	0.12	0.07	0.09	0.12	0.10	0.11	0.12

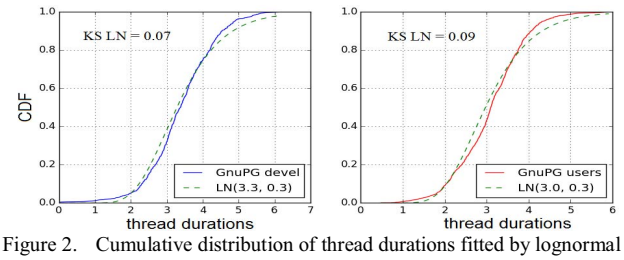

Figure 2. Cumulative distribution of thread durations fitted by lognormal laws in lin-log scale.

Figure 3 shows that for duration $x \geq 720$ minutes i.e. half a day, corresponding to the green vertical line, blue curves are systematically above red curves. This means that the proportion of threads that last more than half a day (i.e. long debates) is greater for the group of developers than for users.

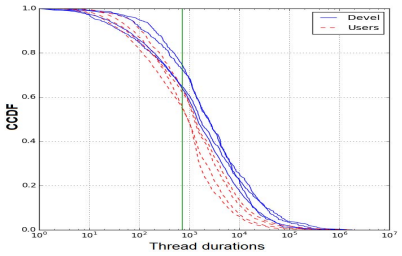

Figure 3. Complementary cumulative distribution of thread durations in for all mailing lists in lin-log scale.

C. Relationship between duration and number of messages

In all mailing lists, discussion threads which have large durations do not have the largest numbers of messages. Such threads can correspond to topics on which members of the mailing list have little to say and that have been abandoned and reactivated later. On the other hand, discussion threads with a large number of messages are short in time. This is probably due to the fact that threads that have the largest number of messages correspond to hot topics related to news, and usually such topics do not last.

These two observations are illustrated in figure 4 by the fact that the upper right corners of the two graphics contain very few points, i.e. thread with very large number of messages and with very large duration are uncommon.

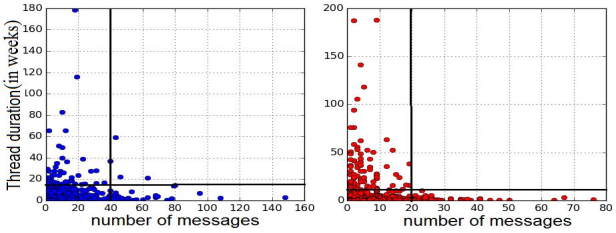

Figure 4. Relationship between thread duration and number of messages. The thread durations are in weeks.

D. Common dominant members

We are interested in the most active participants like in [9]. More precisely, we define dominant members (DM) as the 5% most active participants. The three other categories of participants shown in table IV are members who are registered in the two mailing lists (CM), members who are common to both lists and are dominant in at least one mailing list (CMD1L) and participants who are common and dominant in both mailing lists (CMD2L).

Column 5 in table IV shows that, in each project, the majority of dominant members in developers' mailing list are enrolled in the two mailing lists.

Column 6 shows that roughly half of dominant members in the developers' list are also dominant in the users' list for

CentOS (48%) and GnuPG (56%). For Mailman and Samba, this proportion falls to 38% and 40% respectively. In all projects, the leaders are dominant in both lists.

TABLE IV. DOMINANT GROUPS AND COMMON MEMBERS TO BOTH MAILING LISTS TYPES.

Project	*Type*	*CM*	*DM*	*CMD1L*	*CDM2L*
CentOS	*devel*	240	33	**25 (76%)**	**16 (48%)**
	users	240	136	43 (32%)	16 (12%)
GnuPG	*devel*	104	16	**14 (88%)**	**9 (56%)**
	users	104	70	18 (26%)	9 (13%)
Mailman	*devel*	56	13	**7 (54%)**	5 (38%)
	users	56	65	20 (31%)	5 (8%)
Samba	*devel*	524	86	**61 (71%)**	34 (40%)
	users	524	236	101 (43%)	34 (14%)

E. Event detection

We have analyzed the evolution in time of the number of messages exchanged in each mailing list every 6 months. Our hypothesis is that outstanding events may correspond to sharp changes (sudden increase or decrease) in the slope. It can be seen in figure 5 that for CentOS and Mailman, and to a lesser extent for Samba, events can be detected using the curves of either the developers or the users mailing list. On the contrary, the two curves are quite different for GnuPG. This last observation may be due to the fact that for cryptography, developers are mainly from highly specialized scientific fields while the pool of users is very wide. As a consequence, these two communities are not sensitive to the same events.

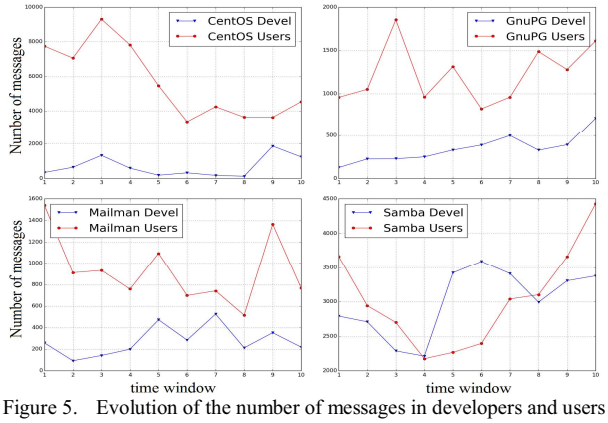

Figure 5. Evolution of the number of messages in developers and users mailing lists. Unit on the x-axis corresponds to 6 months

F. Characteristic inter-link time

We have selected the threads with duration ranging from one day to one week and which contain at least 7 messages. Then we have computed their characteristic inter-link times and plotted for each project, the values obtained for developers and users mailing lists (red and blue colors).

It appears clearly in figure 6 that these two curves are very similar for CentOS, GnuPG and Samba, and are very close for Mailman. Moreover, these curves can be fitted by lognormal functions as shown in figure 6 (green color). This shows that there is no significant difference in the time structures of discussion threads for developers and users mailing lists.

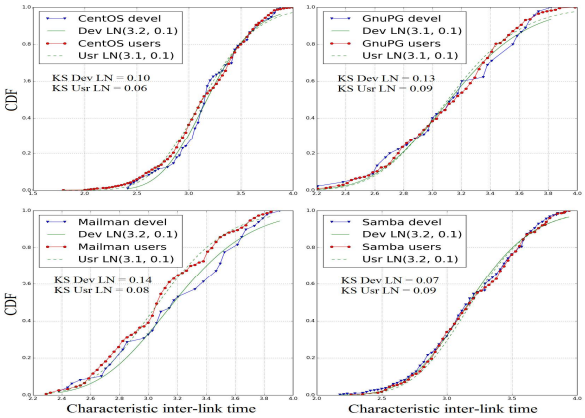

Figure 6. Cumulative distribution of characteristic inter-link time of discussion thread in developers and users mailing lists.

V. RELATED WORK

Sowe et al. [11] have shown that the distribution of the posts and replies of contributors follow power law distributions for both users and developers mailing lists of Debian project. They have also noticed that both variables have largest maximum value, deviation, skeeness and kurtosis in the users list.

Wang [18] has shown that for Debian project, 17% of threads have durations of less than one hour, while 0.3% have durations of more than 115 days. She has also noted that the correlation between size and duration is very weak. Shihab et al. [9] have found that the mailing list activity is driven by a dominant group of participants.

Viard et al. [17] have used the concept of characteristic time to study captures of IP traffic that contain information on very different kinds of activities like file transfers, users interacting with remote systems, automatic backups, or distributed computations.

A distribution which has a characteristic time of the order of half a day is the complementary distribution of inter-contact time in mobility traces [19]. This distribution decays exponentially after the characteristic value.

Kaltenbrunner et al. [20] have shown that a mixture of two log-normal distributions combined with the circadian rhythm of the community is able to explain accurately the reaction time for comments within a discussion thread.

Other studies used mailing lists to describe the structural characteristics of social networks [1, 21, 3, 22].

Lakhani et al. [12] found that at least 50% of public posted questions in Apache Usenet were initially answered on the day of/after posting.

Other studies focus on threads description [6, 1, 22]. In [23], Conein and Latapy model threads as networks and show that threads with line shape structure are those that promote the production of new knowledge. In addition, Delanoë [6] shows that the discussion threads of Debian

mailing lists are moving towards a line shape and explains that this is the result of the control policy of Debian leaders.

Guzzi et al [4] analyzed a sample of 506 email threads from the developers' mailing list of Lucene project. Their work reveals that implementation details are discussed only in about 35% of the threads.

Barcellini et al [5] have compared online interactions for a successful pushed-by-users design process with unsuccessful previous proposals in python project. They found that the cross-participants foster the design process and act as boundary spanners between users' and developers' communities.

VI. CONCLUSION

In this paper, we have made a social network analysis and comparison of developers' and users' mailing lists of four free open source software projects: CentOS, GnuPG, Mailman and Samba.

Our study highlights the usefulness of power law and lognormal law for the analysis of mailing lists features. The duration of 720 minutes or half a day seems to correspond to a characteristic value in a comparative study of developers and users activities.

The GnuPG project seems to have some specific characteristics as compared to CentOS, Mailman and Samba. Indeed, the distribution of messages in threads and its temporal evolution is quite different from what is observed for the three other projects.

These empirical results presented here have permitted us to have a deeper understanding of threads in developers' and users' mailing lists. In future work, we will try to exploit this in order to calibrate prediction models related for instance to the evolution of threads in mailing lists.

VII. ACKNOWLEDGEMENT

This work was supported by the African Center of Excellence in Information and Communication Technologies (CETIC) created by the Cameroonian Government at the University of Yaoundé I, with the support of the World Bank

REFERENCES

[1] R. Dorat, M. Latapy, B. Conein and N. Auray, "Multi-level analysis of an interaction network between individuals in a mailing-list," *In Annales des télécommunications,* vol. 62, pp. 325-349, Springer-Verlag, 2007.

[2] A. Bohn, I. Feinerer, K. Hornik and P. Mair, "Content-based social network analysis of mailing lists," *The R Journal,* vol. 3, pp. 11-18, 2011.

[3] J. Howison, K. Inoue and K. Crowston, "Social dynamics of free and open source team communications," *Open Source Systems,* pp. 319-330, Springer US, 2006.

[4] A. Guzzi, A. Bacchelli, M. Lanza, M. Pinzger and A. van Deursen, "Communication in Open Source Software Development Mailing Lists," *In Proceedings of the 10th Working Conference on Mining Software Repositories,* pp. 277-286, IEEE Press, 2013.

[5] F. Barcellini, F. Détienne and J.-M. Burkhardt, "Cross-Participants: fostering design-use mediation in an Open Source Software community," *In Proceedings of the 14th European conference on Cognitive ergonomics: invent! explore!,* pp. 57-64, ACM, 2007.

[6] A. Delanoë, "Open Source Community Watch Towards a mailing-list socio-meter," *Intelligence Journal,* vol. 2, 2013.

[7] Y. Ye and K. Kishida, "Toward an Understanding of the Motivation of Open Source Software Developers," *Software Engineering, 2003. Proceedings. 25th International Conference on,* pp. 419-429, IEEE, 2003.

[8] D. German, "Using software trails to rebuild the evolution of software," *Journal of Software Maintenance and Evolution: Research and Practice,* vol. 16, pp. 367-384, 2004.

[9] E. Shihab, N. Bettenburg, B. Adams and A. E. Hassan, "On the Central Role of Mailing Lists in Open Source Projects: An Exploratory Study," *New frontiers in artificial intelligence. Springer Berlin Heidelberg, 2010,* pp. 91-103, 2010.

[10] B. Vasilescu, A. Serebrenik, P. Devanbu and V. Filkov, "How Social Q&A Sites are Changing Knowledge Sharing in Open Source Software Communities," *In Proceedings of the 17ACM conference on Computer supported cooperative work & social computing,* pp. 342-354, ACM, 2014.

[11] S. K. Sowe, I. Stamelos and L. Angelis, "Understanding knowledge sharing activities in free/open source software projects: An empirical study," *Journal of Systems and Software,* vol. 81, no. 3, pp. 431-446., 2008.

[12] K. R. Lakhani and E. von Hippel, "How open source software works free user-to-user assistance," *Research policy,* vol. 32, no. 6, pp. 923-943, 2003.

[13] D. D. S. Price, "A general theory of bibliometric and other cumulative advantage processes," *Journal of the American society for Information science,* vol. 27, no. 5, pp. 292-306, 1976.

[14] M. Mitzenmacher, "A Brief History of Generative Models for Power Law and Lognormal Distributions," *Internet mathematics,* vol. I, no. 2, pp. 226-251, 2004.

[15] A.-L. Barabasi and A. Réka, "Emergence of Scaling in Random Networks," *science,* vol. 286, no. 5439, pp. 509-512, 1999.

[16] A. Clauset, C. R. Shalizi and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM review,* vol. 51, pp. 661-703, 2009.

[17] J. Viard and M. Latapy, "Identifying roles in an IP network with temporal and structural density," *Sixth IEEE International Workshop on Network Science for Communication Networks (NetSciCom 2014),* 2014.

[18] Q. Wang, "Link Prediction and Threads in Email Networks," *International Conference on Data Science and Advanced Analytics,* 2014.

[19] T. Karagiannis, J.-Y. Le Boudec and M. Vojnovic, "Power law and exponential decay of inter contact times between mobile devices," *Mobile Computing, IEEE Transactions on,* vol. 9, no. 10, pp. 1377-1390, 2010.

[20] A. Kaltenbrunner, V. Gomez and V. Lopez, "Description and Prediction of Slashdot Activity," *In Web Conference, 2007. LA-WEB 2007. Latin American,* pp. 57-66, IEEE, 2007.

[21] C. Bird, D. Pattison, R. D'Souza, V. Filkov and P. Devanbu, "Latent Social Structure in Open Source Projects," *In Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering,* pp. 24-35, 2008.

[22] V. Gómez, A. Kaltenbrunner and V. López, "Statistical Analysis of the Social Network and Discussion Threads in Slashdot," *Proceedings of the 17th international conference on World Wide Web,* pp. 645-654, 2008.

[23] B. Conein and M. Latapy, "Les usages épistémiques des réseaux de communication électronique: Le cas de l'open-source," *Sociologie du travail,* vol. 50, no. 3, pp. 331-352, 2008.

[24] K. Crowston, K. Wei, J. Howison and A. Wiggins, "Free/Libre open-source software development: What we know and what we do not know," *ACM Computing Surveys (CSUR),* vol. 44, no. 2, p. 7, 2012.