Communities detection and analysis of their dynamics in collaborative networks

Manel BEN JDIDIA
CITI Lab
INSA-Lyon – 69621 Villeurbanne
Email: manel.ben-jdidia@insa-lyon.fr

Céline ROBARDET
LIRIS – UMR5205
INSA-Lyon – 69621 Villeurbanne
Email: celine.robardet@insa-lyon.fr

Eric FLEURY
INRIA ARES - INSA-Lyon – 69621 Villeurbanne
Email: eric.fleury@inria.fr

Abstract

In this paper we propose a new way to identify communities in evolving graphs like collaborative networks. We apply this approach on the Infocom co-authorship network to determine stable collaborations and evolving communities. Finally, we analyse the impact of the co-authorships relation topology on the formation of the program committee board of the conference.

1. Introduction

In a social network, nodes represent social entities (*e.g.*, people) and links indicate social interactions (*e.g.*, friendship, collaboration). Social networks have been central objects of study in the social sciences for a long time, since they have the potential to highlight how social outcomes can arise not just from the properties of individuals in an isolated manner, but from the pattern of interactions among them. More precisely, we are interested by the intrinsic structure of the network, by identifying patterns of contacts or interactions between groups of people. Most of the interesting features of real-world social networks that have attracted the attention of researchers in the last few years reveals that such networks are not like random graphs (first studied by Rapoport [10], Solomonoff and Rapoport [11] and Erdös and Rényi [4]).

Relations between researchers are very dynamic. New links may appear all the time due to the network growth or his change over years. So it is interesting to focus the analyze of social networks to the dynamics of these relations in order to better understand the evolution of the interactions between people [1, 7]. As state above, social networks are non random. This important feature suggests that network structure formation should be investigate. It is also widely assumed that most social networks show "community structure", *i.e.*, groups of vertices that have a high density of edges within them, with a lower density of edges between groups. The study of groups and communities appears to be a key feature in the analysis of phenomena based on sociological data since it may help illuminate how the organization's global decision-making behavior is structured. Understanding the structure and dynamics of social groups is a natural goal for network analysis, since such groups tend to be embedded within larger social network structures.

Our specific contributions in this paper are as follows. First we propose two methods to identify stable collaborations and time evolving communities within a co-authorship network in the section 2. Then, we apply this method to the co-authorship network of one major conference, namely INFOCOM[1] in the section 3. To do this, we propose a complete data set about Infocom. Our database includes 23 conference's years among 26. Finally, to study the Infocom social network, we add the members of Infocom Program Committee boards so we can analyse relations between authors and committee members in the Section 4. Finally, the section 5 offers some concluding remarks.

2. Methods to identify evolving communities

Scientific collaborations are often influenced by demographic locations (*e.g.*, authors may be in the same laboratory, institute, research group), personal choices (*e.g.*, friendship between authors, field of publication) or opportunities. In time, this promotes new collaborations between scientific authors, and changes the behaviour of the co-authorship network structure of a conference or a group of conferences. The nodes of the co-authorship network are the authors and there is an edge between two nodes/authors u and v if u and v have been co-author of at least one paper. Each year, new authors may appear extending the node set of the network and new collaborations may append leading to new links between nodes.

[1] IEEE Communications Society Conference on Computer Communications

The analysis of such a social evolving graph can be achieved by the identification of communities. Community definition is very general and depends on the context. The most employed one is that the community structure represents a group of nodes within the connections are denser than those to the other groups. In a co-authorship network, communities exist further to many collaborations within a group of authors. Such entities allow identifying persons sharing a similar part in the network, which allow to have a global vision of the interactions between persons who are in the same group not only with the neighbors. In scientific collaboration field we can examine the relations between authors, the impact of these relations on the scientific board (*e.g.*, the dominance of certain authors, the program committees board construction).

The community structure brings out much information about the network and raised a great interest the last years, especially the community detection problem in a static network [8, 9]. This notion of community is however difficult to define formally. Most recent approaches have reached a consensus, and consider that a partition $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ of the vertices of a graph $G = (V, E)$ $(i = 1 \ldots k,\ \mathcal{C}_i \subseteq V)$ represents a good community structure if the proportion of edges inside the $\mathcal{C}_i$ (internal edges) is high compared to the proportion of edges between them. For example, [6] et al. propose to use the temporal evolution of the degree as descriptive feature of each node and use the EM clustering algorithm to identify communities. It can be easier to identify communities in a static graph than considering the dynamic aspect of the network actors. Hopcropft et al. [5] propose a first approach in that direction. It consists to first identify *natural communities* as stable clusters obtained using an agglomerative clustering algorithm at different points in time, and then associate to each natural community the best matching natural community in the next time step.

In the following, we propose two approaches to identify communities. In the first one, time stable collaborations are captured thanks to an exhaustive computation approach that consists to define the searched patterns thanks to constraints and to use an algorithm that output all and only all patterns that satisfy the constraints. The second approach consists to identify evolving communities in a single step process based on a random walk clustering approach. This approach enables to define a distance between nodes that is related to the eigenvectors and the eigenvalues adjacency matrix of the graph, known to be important characteristics to identify communities. The random walk computation approach is more computationally efficient than the traditional approach that enable to obtain the spectral properties of the matrix.

2.1. Identifying stable collaborations

Dynamic communities in a co-authorship network can be considered as groups of authors that frequently co-sign papers. Considering the co-author relationship of one year as a graph, a first approach can consist in computing frequent maximal cliques in this set of graphs. More formally, let us define a graph G_t as the co-author graph for year t. We have $G_t = (V_t, E_t)$ with V_t the set of authors that published a paper during year t, and E_t is the set of edges $\{x, y\}$ such that $x, y \in E_t$ and, x and y co-sign a paper that was published the year t. A maximal clique C on G_t is such that $C \subseteq V_t\ such\ that\ \forall x, y \in C,\ \{x, y\} \in E_t$ and $\forall z \in V_t \setminus C,\ \exists x \in C\ such\ that\ \{x, z\} \notin E_t$ The size of C is its cardinal $|C|$.

Let us now consider a set of T such graphs, i.e. $\mathcal{G} = \{G_1, \cdots, G_t, \cdots, G_T\}$. A clique C in $\mathcal{G}$ is such that $C \subseteq \bigcup_{t=1}^{T} V_t$ and its associated set of graphs S that supports it is defined by $S = \{t \mid \forall x, y \in C,\ \{x, y\} \in E_t\}$. We say that a such clique is frequent w.r.t. σ if $|S| \geq \sigma$ and maximal if $\forall z \in \bigcup_{t=1}^{T} V_t \setminus C,\ \exists x \in C\ and\ \exists s \in S\ such\ that\ \{x, z\} \notin E_s$. Efficient algorithms [3, 2] enable to compute frequent maximal cliques in large evolving graph and thus enable the identification of stable collaborations in the co-authorship network.

2.2. Identifying evolving communities

To discover evolving communities, we adapt the random walk based method of Pascal Pons to evolving graphs. Pascal Pons [9] proposes to identify social communities in a single graph $G = (V, E)$ using random walk. Such communities are defined as dense area in the graph, where vertices of a community are strongly connected, whereas they have fewer links towards outside. The main idea is that random walks would be trap into dense area thanks to the high density of links in the community. Using a short distance walk (let say 4 steps), a walker may stay in its original community. A random walk from a vertex defines a vector of probabilities to reach others vertices. Comparing the vectors associated to two distinct vertices enable to evaluate their proximity. The identification of communities is done by the algorithm WALKTRAP which computing a hierarchical clustering algorithm on this similarity matrix and then selecting the partition P that maximizes the coefficient of modularity. Each cluster of the partition is thus considered as a community. The coefficient of modularity of a partition P is defined by: $Q(P) = \sum_{C \in P} e(C) - a(C)^2$ where $e(C)$ is the proportion of edges that are intern of the cluster C, i.e. $e(C) = \frac{1}{2\,|E|} |\{\{x, y\} \in E\ such\ that\ x \in C\ and\ y \in C\}|$. $a(C)$ is the number of edges that are linked to the community: $a(C) = \frac{1}{2} |\{\{x, y\} \in E \mid x \in C\ and\ y \in V\}|$. In a random graph, the expected proportion of intern links is equal to $a(C)^2$ and thus Q compares the effective proportion of intern links to the expected one on a random graph.

Here we propose to use this approach to analyze dynamic social communities. To capture the proximity between authors using random walk, we propose to view the evolving network as a single evolving graph $G_{evol} = (V, E)$ where E is a set of author-year pairs and E is defined as follows: *(i)* there is an edge between the node $[i, t]$ (the author i at time t) and the node $[j, t+1]$ (the authors j at time $t+1$) if there exists an author k such that i and k are co-authors at time t and, k and j are co-author at time $t+1$. These edges are called *transversal edges*; *(ii)* there is an edge between $[i, t]$ and $[j, t]$ if i and j are co-authors at time t; *(iii)* there is an edge between $[i, t]$ and $[i, t+1]$ if i has published a paper at year t and $t+1$.

Fig. 1 illustrates the graph construction. This graph links authors that are co-authors but also authors that have a common co-author in one year of interval. For example, authors 1 and 2 co-sign a paper at year $t-1$ and authors 2 and 7 co-sign another paper at year t. Thus there is a transversal edge between $(1, t-1)$ and $(7, t)$. We do that because it is usual to continue collaboration with co-authors even if it does not leads to a publication. It can also append that a co-author finishes a work alone, but the two authors still belong to the same community. We will see in the following that removing such links leads to communities of worst quality.

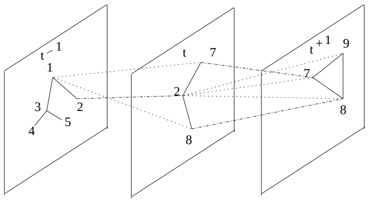


Figure 1. An example of graph.

3. Infocom communities

In order to apply the proposed method to a co-authorship network, we consider the co-authorship network of the Infocom conference. First, we collected the Infocom data from the period between 1985 and 2007. For each year, we gather the papers published and their authors. The main source for our data was DBLP[2] where we found a lot of paper titles and the authors' names and we got miscellaneous data from INRIA document centers by ordering facsimile of oldest data that were not online. An important work has gone into disambiguation of similar names, so co-authorship relationships are relatively free of name resolution problems.

Based on the information gathering, we were able to build a database of 4030 unique papers from 1985 to 2007. The database contains also 5164 unique authors over a 23 years time period. Table 1 describes the general characteristics of the co-authorship network of Infocom. We also add in Table 1 the same general characteristics for the co-authorship network of arXiv[3]. The arXiv graph is larger than the Infocom graph (3 times), but many properties of both graphs are close. The average degree k of all nodes of both co-authorship networks is approximately 3.5 (*i.e.*, on average, an author has 3.5 co-authors). We also have, in both cases, an average distance d close to 7. We can notice here one property of the *small world* phenomena where the number of nodes n is large, the graph is sparse (m is roughly linear in n) whereas the distance between two nodes is relatively small. Thus, Infocom data is quite representative of general co-authorship networks.

	arXiv	**Infocom**
n (nb. of vertices)	16 401	5 164
m (nb. of edges)	29 552	8 918
k (average degree)	3.60	3.45
δ (density)	2.2e-4	6.6e-4
d (average distance)	7.18	6.92
diameter	20	18

Table 1. General characteristics of co-authorship network.

To identify stable collaborations in the Infocom co-authorship network, we first construct the co-authorship network G_{evol} from 1985 to 2007. We use DATA-PEELER [2] to compute maximal frequent cliques. In this algorithm, in addition to the frequency, man can also use a minimal size constraint on the sizes of the computed cliques. Thus the algorithm enables to compute only large and frequent cliques, which are the most interesting ones. We run DATA-PEELER on the $T = 23$ co-author graphs of Infocom enforcing the clique size to be above or equal to 3 and the frequency σ was also fixed to 3. We obtain the 25 cliques presented in table 2. By analyzing the groups which we found, we can see that in 15 cases, people who form the groups are authors having the same order of degree. But in other cases we find authors with a high degree who publish with authors having a much lower degree. We find in these cases that the authors are a part of the same research group which support their many collaborations.

To identify evolving communities, we construct the graph G_{evol} on the Infocom co-authorship data. The obtained graph contains 8 692 vertices and 29 972 edges. We apply the WALKTRAP algorithm on it and we obtained 90 communities having a Maximal modularity Q of 0.502547. Fig 2 (top) shows the distribution of the sizes of these communities : half of them are quite small, but few of them are quite large. If we remove the transversal edges, the graph has 8 692 vertices and 12 467 edges, i.e. more than the half of edges of the previous graph are transversal edges. On this graph, WALKTRAP computes 126 communities with $Q = 0.366183$, and thus the quality of these communities is below the one obtained on the graph with transversal edges. If we compare the distribution of the sizes of the two sets of communities (see Fig 2 top and bottom), we can observe that transversal edges prevent the computation of lots

[2] DBLP: Digital Bibliography & Library Project, `http://dblp.uni-trier.de/`

[3] `arXiv.org` is an open access site to 427136 electronic prints in Physics, Mathematics, Computer Science and Quantitative Biology

#	Author cliques C				Associated set S of Years				
1	M. Ahamad (2)	M. Ammar (34)	J. Bernabeu-Auban (2)		1988	1989	1990		
2	C. Barnhart (2)	A. Ephremides (8)	J. Wieselthier (6)		1991	1993	1994		
3	D. Dutta (2)	A. Goel (19)	J. Heidemann (15)		2003	2004	2005		
4	L. Kalampoukas (4)	K. Ramakrishnan (23)	A. Varma (2)		1997	1998	2000		
5	R. Doverspike (7)	G. Li (6)	D. Wang (5)		2002	2006	2007		
6	M. Conti (4)	E. Gregori (4)	L. Lenzini (8)		1990	1992	1993		
7	S. Acharya (3)	B. Gupta (4)	P. Risbood (4)	A. Srivastava (4)	2003	2004	2005		
8	M. Kodialam (16)	T. Lakshman (26)	S. Sengupta (6)		2004	2005	2006		
9	S. Low (30)	A. Tang (5)	J. Wang (8)		2003	2004	2005		
10	S. Donatelli (4)	M. Marsan (24)	F. Neri (18)		1990	1991	1992	1993	
11	D. Figueiredo (7)	J. Kurose (53)	D. Towsley (88)		2001	2003	2006		
12	O. Frieder (6)	X. Li (10)	P. Wan (18)		2000	2001	2004		
13	Q. Fang (6)	J. Gao (10)	L. Guibas (6)		2004	2005	2006	2007	
14	M. Azizoglu (8)	A. Somani (9)	S. Subramaniam (10)		1996	1997	1998		
15	G. Iannaccone (7)	S. Jaiswal (9)	J. Kurose (53)	D. Towsley (88)	2003	2004	2006		
16	Y. Breitbart (9)	M. Garofalakis (14)	R. Rastogi (18)		2000	2001	2002	2003	
17	C. Hollot (9)	V. Misra (16)	D. Towsley (88)		2001	2002	2003		
18	I. Cidon (24)	A. Khamisy (7)	M. Sidi (26)		1992	1993	1994	1997	1998
19	A. Bianco (7)	E. Leonardi (18)	M. Marsan (24)	F. Neri (18)	1993	1996	1997	2001	
20	A. Bianco (7)	E. Leonardi (18)	F. Neri (18)		1993	1996	1997	2001	2004
21	S. Bhattacharjee (18)	K. Calvert (10)	E. Zegura (19)		1996	1998	2000		
22	P. Giaccone (9)	E. Leonardi (18)	F. Neri (18)		2001	2003	2004		
23	P. Giaccone (9)	E. Leonardi (18)	M. Marsan (24)		2001	2003	2004		
24	E. Leonardi (18)	M. Marsan (24)	M. Mellia (12)	F. Neri (18)	2000	2001	2002	2005	
25	E. Leonardi (18)	M. Marsan (24)	F. Neri (18)		1993	1996	1997	2000	
					2001	2002	2003	2005	

Table 2. The 25 cliques (the degree of each author is given between brackets)

of small communities of few interest. The transversal edges enable the computation of larger communities.

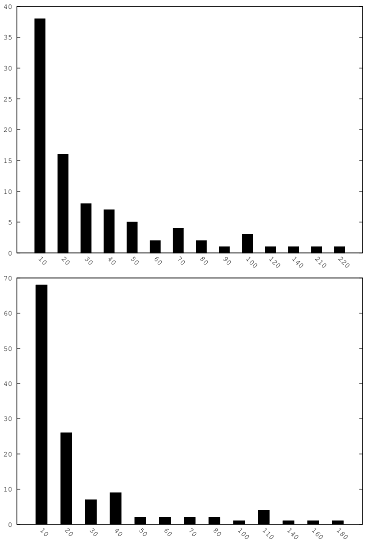


Figure 2. Distribution of the sizes of the communities (top) with transversal edges, (bottom) without.

To have a closer look to the obtained communities on the graph with transversal edges, we consider those that extend one of the previous cliques (the clique 13). Fig 3 shows the dynamic community that corresponds to the clique 13. The community gathers the three previously identified frequent co-authors (Q. Fang, J. Gao and L. Guibas) but also other authors that have been identified as belonging to the same community. All these authors are co-author which was not mandatory due to the way we constructed the graph. The way communities are computed support dense sub-graphs, i.e. co-author links play an important role.

Some of the cliques have no corresponding community in the WALKTRAP approach. For example, the second clique is one of them. It can append for not temporally continuous one (e.g. clique numbers 2, 5, 6, 12, 19, 20, 25). This clearly due to our graph definition that considers only edges between t and $t+1$. But some of the not temporally continuous cliques have been successfully identified as communities: cliques 4, 11, 15, 18, 21. We can notice that cliques 4 and 11 are not identified as communities in the graph without transversal links. Cliques numbered 21, 22 and 23 have been merge in a single community.

4. Impact of the co-authorship topology structure on program committee boards

The fact of joining the Infocom PC can be considered as a mechanism of diffusion of innovation [1] such that PC members are co-opted by one of their relations/colleagues. The underlying hypothesis in diffusion studies is that individual's probabilities of adopting a new behavior increases with the number of friends already in the community. Figure 4 shows the proportion $P(k)$ of authors who join the Infocom PC as a function of the number k of their co-authors who are already member of the PC for three different years.

We can observe that globally there is a positive correlation between writing a paper with PC members and entering the program committee. The decreases of the curves and irregularities are mainly due to the small number of data.

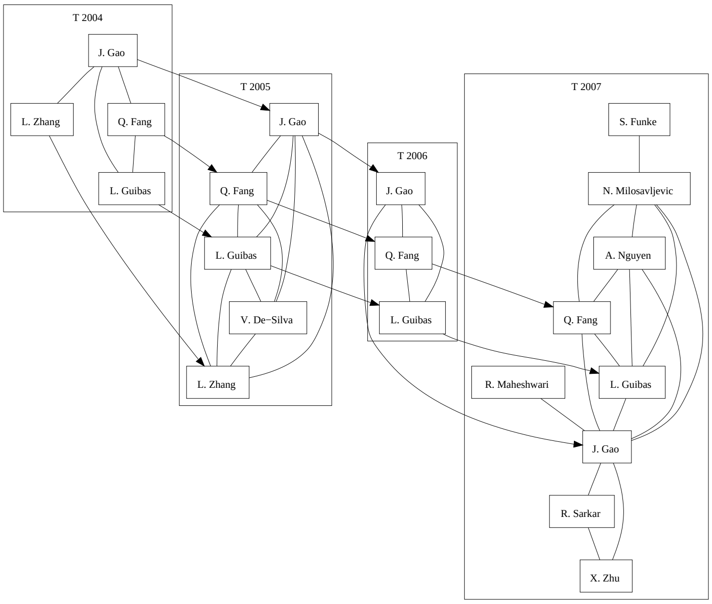


Figure 3. Community that corresponds to the clique 13 of table 2.

Note that if we increase the co-authoring relation up to distance 2 in order to increase the number of data (*i.e.* by considering co-authors of co-authors), the correlations is not relevant any more. The fact is that the number of co-authors in the PC has a positive impact even if it is not preponderant but the relation between co-authors does not impact so much.

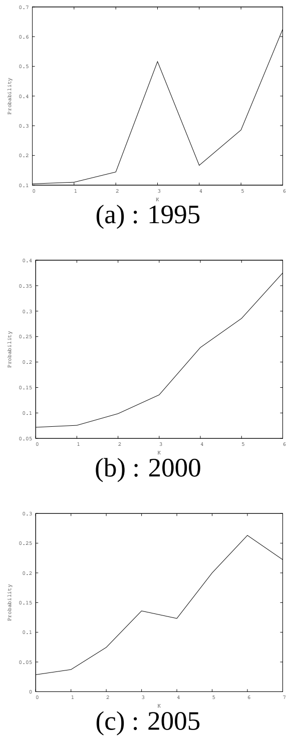
To have a closer look at the dependencies that may exist between the co-authorship graph topology and the PC membership property, we follow the methodology described in [1]. It consists of using features that describe the structure of the graph to construct a predictive model of PC membership property. To make estimates about joining the Infocom PC, we compute a decision tree based on the features described in table 3. A decision tree is a tree-structured plan of a set of features to test in order to predict the output. In our context, we want to construct a model that enables to predict the PC member property. To decide which feature should be tested first, one can simply find the one with the highest information gain. Our goal is to describe the link that may exist between the features listed above and the PC property. We will not use it for a prediction purpose.

We obtain the tree shown in Figure 5. This tree enables us to characterize with a high level of precision authors that are not member of the PC. Among the 4441 authors of this class, 4350 are correctly classified. PC member are more difficult to characterize: the tree manages to correctly classify 363 PC authors among the 723 individuals of this class. Currently, the large majority (341 of 360) of misclassified individuals of this class has less than three publications in the Infocom conference and thus there are not enough topological elements that allow us to characterize them. This is also the case for the additional 232 PC members that had never published in Infocom during this period.

Feature number	Feature description
1	Connected component size
2	Number of co-authors (degree)
3	Number of years where an author has published at least one paper
4	Number of co-authors of a given author that are also co-authors together (similar to the clustering coefficient)
5	Proportion of co-authors of a given author that are also co-authors together
6	Number of co-authors that are PC members
7	Number of published papers
8	Number of co-authors at distance at most 2
9	Number of co-authors at distance at most 3

Table 3. Features related to an author used by the predictive model.

In the tree, one may remark that PC members have more than 3 published papers. It also appears that PC members have published more than 4 times at Infocom, or they have more than 4 co-authors that are PC members. Another way to characterize PC members is to remark that they have a large proportion of co-author that are also adjacent co-authors and they belong to connected component of large size (greater than 8). Finally, the last case is when the number of co-authors at distance at most 2 is important (greater than 33). If the number of co-authors that are also PC mem-



(a) : 1995

(b) : 2000

(c) : 2005

Figure 4. Prob. of entering the PC w.r.t. the nb. of co-authors that are already in the PC

bers seem to be an important feature, one may also note that the structure between co-authors of a given author play an important role. It seems that having co-authors that are densely connected impact on the PC member property.

5. Conclusion

In this paper we emphasized the importance of detecting and investigating the communities in a dynamic collaborative network. We propose two methods that enable the identification of stable collaborations and evolving communities while taking into account their temporal evolutions. We apply these approach on the co-authorship network of the Infocom conference where authors represent nodes and co-authoring a paper represent link between paper authors. We also examine the influence of the co-authorship structure on the PC board formation of the same conference. We show that direct co-authors in the Infocom co-authorship network have a significant impact on the PC board.

Understanding the structure and dynamics of social groups is a natural goal for network analysis, since such groups tend to be embedded within larger social network structures. In order to better discern the full dynamics of all scientific authors several works are on going and remain for future investigations. A first on going work is to investigate more about communities embedded within larger social network structures, we plan to extract larger data (from arXiv and DBLP) and see if the Infocom communities remain the same in a larger context. Another investigation is to study the influence of PC members viewed as an Infocom community. Also we plan to gather PC board information of some other famous conferences (*e.g.*, ACM SigCOMM) and determine if one can found a small world structure between different boards such that some communities are able to sit in different boards and gain a strong influence on the whole scientific domain.

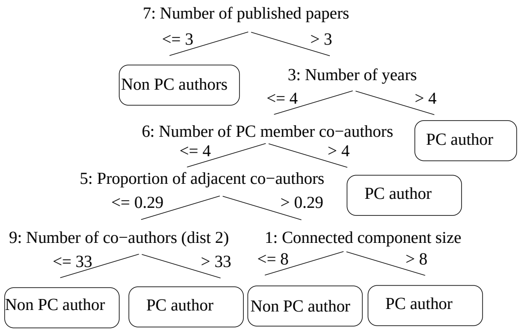

7: Number of published papers
<= 3
> 3
Non PC authors
3: Number of years
<= 4
> 4
6: Number of PC member co−authors
<= 4
> 4
PC author
5: Proportion of adjacent co−authors
<= 0.29
> 0.29
PC author
9: Number of co−authors (dist 2)
<= 33
> 33
1: Connected component size
<= 8
> 8
Non PC author
PC author
Non PC author
PC author

Figure 5. Decision tree

References

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06*. ACM.

[2] J. Besson, L. Cerf, C. Robardet, and J.-F. Boulicaut. Data-peeler. Research report, 2007.

[3] E. Davies. The minimal match graph and its use to speed identification of maximal cliques. *Signal Processing*, 22(3):329–343, 1991.

[4] P. Erdös and A. Rényi. On random graphs. *Publ. Math. Debrecen*, 6:290–297, 1959.

[5] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *PNAS*, 101:5249–5253, 2004.

[6] E. A. Leicht, G. Clarkson, K. Shedden, and M. E. J. Newman. Large-scale structure of time evolving citation networks. *Physics and Society*, pages 1–10, 2007.

[7] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03*, pages 556–559, 2003.

[8] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69, 2004.

[9] P. Pons. *Détection de communautés dans les grands graphes de terrain*. PhD thesis, Univ. Paris 7, 2007.

[10] A. Rapoport. Contribution to the theory of random and biased nets. *Bull. Math. Biophys.*, 19:257–277, 1957.

[11] S. Solomonoff and A. Rapoport. Connectivity of random nets. *Bull. Math. Biophys.*, 13:107–117, 1951.