Multi-layered Social Network Creation Based on Bibliographic Data

Przemysław Kazienko[1],Piotr Bródka[1,4], Katarzyna Musial[2,3], Jarosław Gaworecki[4]
[1] *Institute of Informatics, Wrocław University of Technology, Wrocław, Poland*
[2] *School of Design, Engineering & Computing, Bournemouth University, Poole, United Kingdom*
[3] *Telnet Sp. z o.o., Wrocław, Poland*
[4] *Research & Engineering Center Sp. z o.o., Wrocław, Poland*
kazienko@pwr.wroc.pl, piotr.brodka@pwr.wroc.pl, kmusial@bournemouth.ac.uk, jaroslaw.gaworecki@rec-global.com

Abstract

A method for extraction of the multi-layered social network based on the data about human collaborative achievements, in particular scientific papers, is presented in the paper. The objects linking people form a hierarchy, which is flattened in the pre-processing stage. Only one level of the hierarchy remains together with new activities moved from its other levels. Separate layers of the final multi-layered social network are created based on these pre-processed activities.

1. Introduction

There are plenty of possibilities for people to interact and collaborate each other using different IT systems. Many of these activities leave their distinctive tracks in the form of different data sets. This data about human interactions and common achievements provides the opportunity to extract social networks existing between people. However, due to their scale, complexity and dynamics, these networks are difficult (or even impossible) to be analysed by means of of traditional social network analysis (SNA) methods. Overall, the relations between users can be extracted based on both direct communication and indirect meetings at common activities. The former is e.g. a teleconference via VoIP, whereas to the latter we can count co-authorship of scientific papers. In both situations there exists an object (e.g. a teleconference, an article) that serves as an medium in linking people. Having these objects together with user activities towards them, different types (layers) of relations can be extracted. Additionally, there may exists some relations between objects themselves, e.g. a book co-edited by some scientists contains some chapters worked out by other authors or is cited by another book.

The extraction of the multi-layered social network in the environment where users diversely co-operate each other by means of certain objects, which, in turn, create a hierarchy, is the main contribution of this paper.

The concept of social network has been described by different researchers [15, 17]. The general definition is that *a social network* is a finite set of individuals, who are the nodes of the network, and activities or relations between them, which represent edges of the network. A social network (SN) commonly represents the mutual communication and activity occurring between users as well as their direction, intensity, and even profile.

During analysis of such networks, researchers usually take into account only one type of activity while in most cases, there are many different relationships between users. The special type of social networks that allows the presentation of many different activities is called a multi-layered social network [9, 17]. Overall, due to their complexity, such networks are more difficult to be extracted and analyzed than simple one-layered social networks.

Social networks extracted from activity data gathered in computer systems were investigated by many researchers; SNs can be extracted from e.g.: bibliographic data [7], blogs [1], photos sharing systems like Flickr [10], email systems [16], telecommunication data [2, 11], social services like Twitter [8] or Facebook [6], video sharing systems like YouTube [5], Wikipedia [4] and much more. Moreover, the whole separate systems were created only for the extraction, aggregation and visualization of social networks [13,14].

Only few scientists have focused their research interests at multi-layer social network extraction from activity data. Moreover, even if there is a little work on this subject no one has studied the hierarchy and relationships between objects in this data. The only hierarchical dependencies in the social networks that were analyzed were associated with the hierarchy between users such as employee-employer, the employee-manager, etc. [12]. Thus, the analysis of

hierarchies between objects is a completely new approach in extraction of multi-layer social networks.

To perform entire extraction process, the hierarchical pre-social network (HPSN), where relation between users and objects as well as between objects exist, must be created. Afterwards, the flattening process, in which the hierarchy of objects is removed, is performed (Section 4). As a result, the flat pre-social network (FPSN) is obtained where the only activities towards one type of the previously chosen objects exist. Based on FPSN, a multi-layered social network with multiple connections between users is created (Section 5). The whole idea is presented at the toy example of bibliographical data (Section 6). Finally, the real-world experiments were performed and their outcomes are presented in Section 7.

2. Hierarchies Between Objects

In all computer-based social networks, which have been analysed in the literature so far, the relationships between users were extracted mainly based on a given type of communication or common activity. For example, if two users exchange emails then the relationship between them in the social network may be established. However, both user communication and common activities are always related somehow with the objects which serve as a medium in interactions between users and their common activities, see section 3. This object may be "a message" in the case of email exchange, "a topic" in the Internet forums or "a scientific paper" co-authored by some researchers. Humans may act towards an object in different ways, e.g. a scientist may either only edit a book or be an author. Additionally, objects can be in hierarchical relationships, e.g. journal *Communication of the ACM* has many issues that contain articles cited by other papers, Fig. 1. Instead of journal we can also have conference proceedings or edited books.

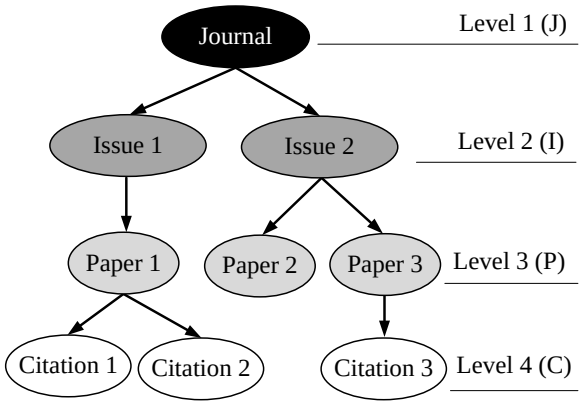

Fig. 1 Hierarchies between objects within bibliographical data

People meet each other by performing activities to objects that belong (i) to one specific level in the object hierarchy (many users can co-author a book) or (ii) to two different levels of this hierarchy , e.g. editor of the conference proceedings is in the indirect relation with authors of the papers published there as well as with authors of citations of these papers.

3. Hierarchical Pre-Social Network

The first step in the social network pre-processing method is cleansing the gathered activity data and creation of the hierarchical pre-social network (HPSN). From obtained data such elements as: users, objects, hierarchy between objects as well as users activities towards objects are extracted. HPSN contains the information about the activities between users and objects towards which they performed some activities. The main feature that is characteristic for HPSN is that there exist hierarchies between different objects. The whole process of HPSN extraction can be divided into four steps.

1) User extraction – the extracted users are the nodes both in the pre-social network and the final social network. Users are people who perform different activities towards various types of objects, e.g. they send emails to each other or comment the photos uploaded by other people. These activities are the basis to create the roles of users in relation to objects, e.g. author, commentator, etc.

2) Object extraction – objects are the nodes in both hierarchical and flat pre-social network, i.e. elements through which users communicate with each other (e.g. email, phone call) or items towards which users perform some activities (e.g. photo, video, tag).

3) Extraction of the hierarchy between the objects – some objects can be in hierarchical relation with other objects, e.g. publications in the bibliography dataset

contains the citations and because a citation cannot exist without publication then they are on the lower level of the hierarchy. The hierarchy causes that the objects on the lower level cannot exist without objects on the upper level. These hierarchies exist within HPSN and are removed during the pre-network flattening process, see Section 4.

4) Extraction of the roles of users towards objects – role of a given user *a* towards an object *X* exists if user *a* performed some activities towards object *X*, e.g. user *a* authored paper *X*. The type of activity that a user performed towards an object is expressed in the name of the role, e.g. if user *a* edited article *X* then the name of the role is *Is Editor*.

The concept of HPSN is presented in Fig. 2 where the hierarchy between objects has got three levels (*A*, *B* and *C*), at each level some objects exist (e.g. at the level *A* objects *OB A1* and *OB A2*) towards which users (*a*, *b*, *c*) perform some activities that result in different types of roles of these users to objects (*X*, *Z*, *V* and *Y* roles), i.e. users performed some activities towards these objects.

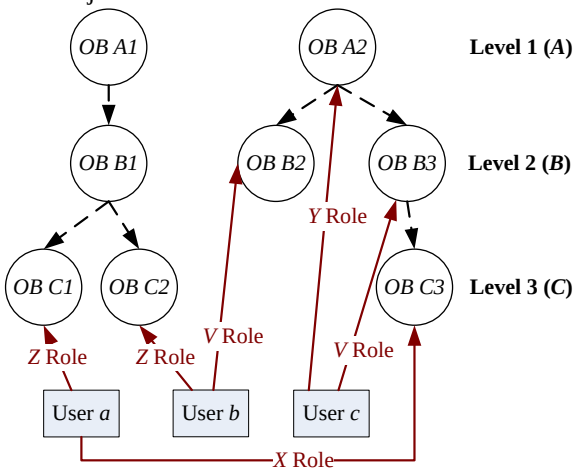

Fig. 2 The concept of HPSN

4. Flat Pre-Social Network

The flattening process aims to remove relationships between objects (hierarchies) and the result of it is the creation of flat pre-social network (FPSN). As a result of this process new types of activities and in consequences roles of users towards objects, which can be extracted from the knowledge about hierarchies, are created.

First step of this process is to decide the level in the hierarchy to which the flattening process will be performed – from now on we refer to it as to the end level. Note that after each flattening process, the only objects in FPSN will be the ones that are on the end level and all users will be connected only towards these objects.

Depending on the position of the end level in the hierarchy three different types of flattening process can be distinguished: (i) bottom-top, (ii) top-bottom, and (iii) mixed. The first one is when the end level is the highest one in the hierarchy and the second one is applied when the end level is the lowest in the hierarchy. In all other cases the mixed flattening process is performed. Below we present in details the bottom-top and top-bottom approaches. The mixed approach is a simple combination of two enumerated above, thus it will not be described.

4.1 Bottom-top flattening process

In the bottom-top approach the roles of people towards objects existing on the hierarchy levels that are below the end level are changed (i.e. User *a* – *OB C1*, User *a* – *OB C3*, User *b* – *OB C2*, User *b* – *OB B2* and User *c* – *OB B3* connections in Fig. 3). The activity from a user to an object from the lower lever is moved to the upper level by:

a) identification of an object on the upper level that is "a father" of the object from the lower level ("child");
b) creation of a new role of the user towards "the father" object;
c) the name of the role of user toward "father" object is created by adding to the name of the role user – "child" the word that denotes the movement from the lower level. For example: in the case of the activity from user *a* to object *OB C1* ("child") the role name is *Z Role* and the name of the new role User *a* – *OB C1* ("father") is ***ZB** Role*;
d) deletion of the activity from the user to the "child" object from the lower level.

NOTE: This process is repeated for other upper levels until the end level is reached (Fig. 3).

The roles of people towards objects existing at the end level remain unchanged (*Y* Role of User *c* toward *OB A2*).

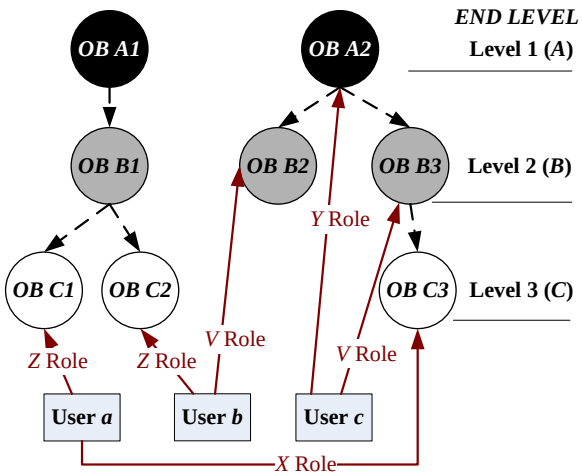

Fig. 3 The hierarchical pre-social network HPSN

The final FPSN in the case of bottom-up approach where the HPSN from Fig. 3 is flatten to level 1 (*A*) is presented in Fig. 4.

4.2 Top-bottom flattening process

In the case of bottom-top approach the roles of people towards objects existing on the hierarchy levels that are above the end level (User *b–OB B2*, User *c–OB B3* and User *c–OB A2* in Fig. 5) are changed.

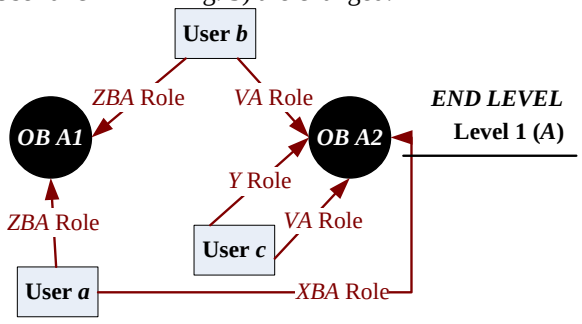

Fig. 4 Structure of the final FPSN after the bottom-top flattening process

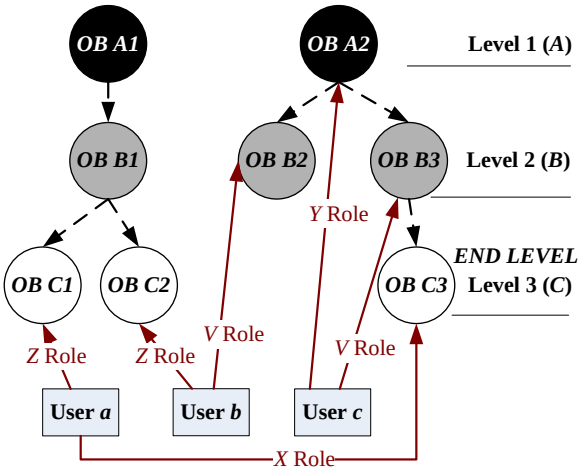

Fig. 5 Example of the structure of the hierarchical pre-social network HPSN

The activity from user toward object from the upper lever is moved to the lower level by

a) identification of all objects on the lower level that are "children" of an object form the upper level ("father"),
b) creation of the new role of the user to all "child objects",
c) the name of the role of the user toward "child object" is created by adding to the origin name information about the "child object". For the example in Fig. 5: in the case of the activity from User *c* to *OB B3* ("father") the role name was: *V* Role and the new role User *c–OB C3* ("child") will have the name: *VC* Role,
d) deletion of the activity from the user to "father object" on the upper level.
e) This process is repeated until the end level is reached (Fig. 6Fig.).

The roles of people toward objects existing on the end level remain unchanged.

In the case of the top-bottom approach, the origin HPSN from Fig. 5 is flatten to level 3 (*Z*), i.e. to FPSN in Fig. 6.

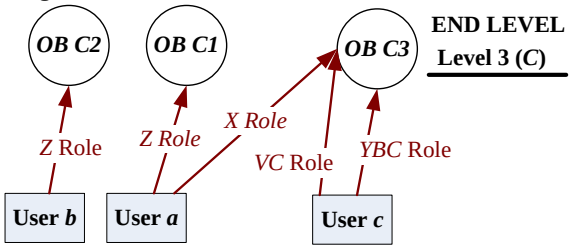

Fig. 6 Structure of the final FPSN after the top-bottom flattening process

5. Social Network

The flat pre-social network structure FPSN is used to extract the social network SN where the activities user-object from FPSN do not exist any more. These connections are converted into direct relations between users in SN. The process consists of the following steps:

a) Extraction of SN layers based on the type of the roles that users have towards objects. Each network layer consists of users and the connections between them.
b) Extraction of relations: *user_from–user_to* by calculation of the relationship strengths and colours between SN nodes (users) using activity data stored in FPSN. The relationship strengths can be calculated differently for different types of relations. There are many possible formulas for this purpose; most are based on the normalized quantity of shared user activities towards objects in FPSN, see [10] for examples.

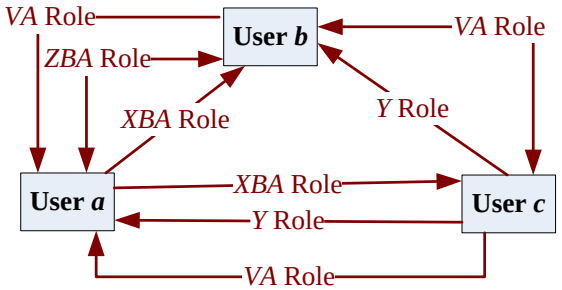

Fig. 7 The social network created from FPSN presented in Fig. 4

The social network SN created from FPSN from Fig. 4 according the process described is presented in Fig. 7 and the SN created from FPSN from Fig. 6 is pictured in Fig. 8. Both networks are multi-layered ones as they consist of more than one type of relationships between users.

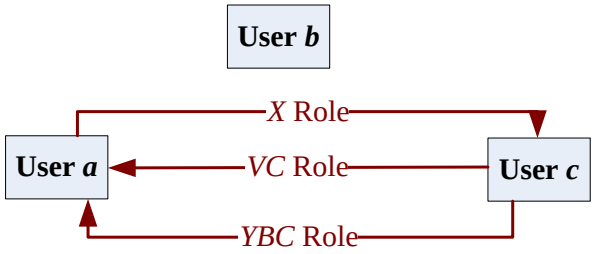

Fig. 8 The social network created from FPSN presented in Fig. 6

6. Example of Flattening Process

The example of a simple hierarchy between object is the bibliographic data. Each publication have a number of references, which are in fact also a publications with references. The hierarchy between objects and also activities that can be performed towards these objects are presented in Fig. 9.

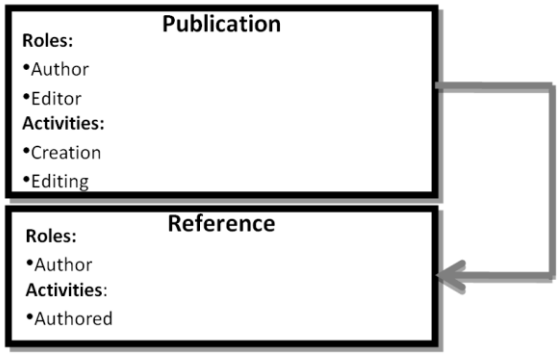

Fig. 9 Objects hierarchy in the bibliographic data

The example of hierarchy between objects and activities between users and objects in the source hierarchical pre-network are presented in Fig. 10. Both, top-bottom and bottom-top flattening processes that aim to create two different flat pre social networks will be described below (See Section 5).

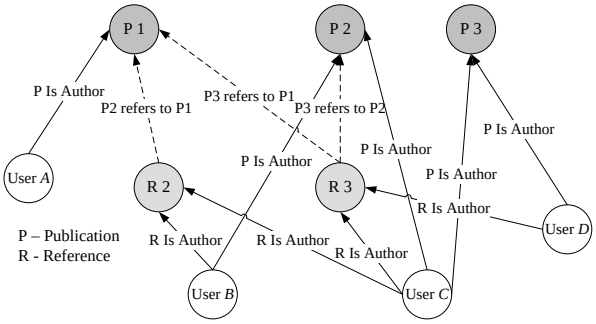

Fig. 10 Relations between users and objects together with users' roles towards the objects

In the bottom-top approach the activities will be moved to the publication level – end level for flattening process. The publication will be the final level and in FPSN all users will be in relations with only publication object. The result is presented on Fig. 11.

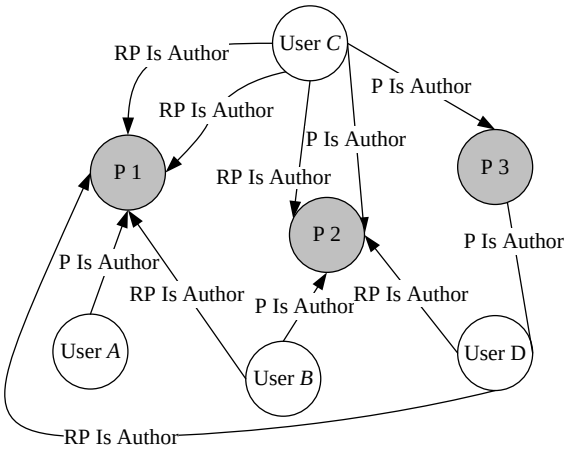

Fig. 11 FPSN after bottom-top flattening process

After applying bottom-top flattening process to HPSN from Fig. 9 new activities emerge:

- User *B* is an author of Reference 2 (R2). A new role of User *B* toward Publication 1 (P1) is formed: *RP Is Author* (*RP – ReferencePublication* Role, see Section 4), i.e. User *B* is an author of R2 and in the same time R2 is cited in P1.
- User *C* is an author of the Reference 2 (R2) that is cited in P1 and Reference 3 (R3) that is cited in P1 and P3. Thus, for User *C* three new activities *RP Is Used* have been created – one towards object P2 and two towards object P1.
- User *D* is an author of Reference 3 (R3) which is cited in P1 and P2. This results in creating two new activities *RP Is Used* of User *D* towards objects P1 and P2..

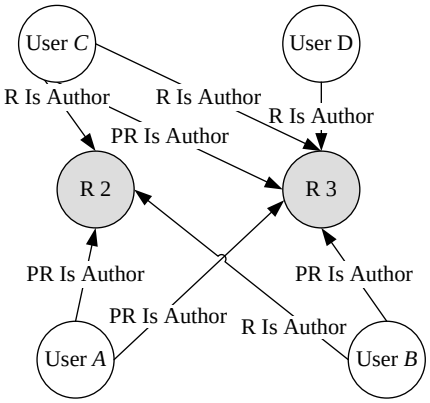

Fig. 12 FPSN after top-bottom flattening process

In the top-bottom method, the activities will be moved to the references level (Fig. 12).

The activities between users and references remain unchanged. However, some new activities are created:

- User *A* as an author of P1 which cites R2 and R3 has a two new roles *PR Is Author* – one toward object R2 and one to object R3.
- User *B* and User *C* are authors of P2 which includes citation to R3. Thus they both received new role *PR Is Author*.

Note, that due to fact that Publication 3 was not citied, during flattening process all roles associated with the Publication 3 has been lost.

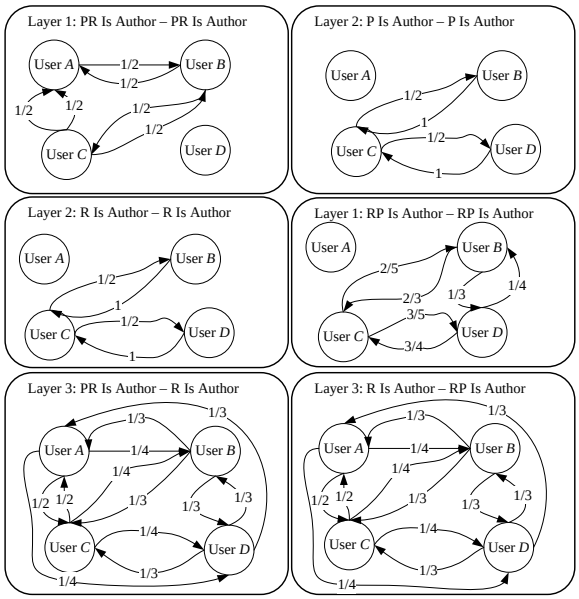

Fig. 13 Three layers of SN extracted from the flat pre-social network FPSN (on the left is top-bottom FPSN, on the right is bottom-top FPSN)

Based on the flat pre-social network FPSN various layers in the social network SN can be identified. Each layer is extracted based on *one* or *two* types of roles that users have towards the objects. From both flat pre-social networks (Fig. 11 and Fig. 12) three different layers can be extracted. All layers are presented in Fig. 13.

Note that User A is not isolated only because of flattening process. Additionally, new relation between User *B* and User *D* was created (Fig 13).

7. Experiments

The *DBLP* Computer Science Bibliography dataset (http://www.informatik.uni-trier.de/~ley/db/) was used for the experiments. The datasets time range was between 2002-01-03 and 2010-02-19. First stage of data processing was cleansing and validation phase. The most important rules that bibliography data has to meet are:

- Each object must have a creation date
- Each object must have an author
- Each object must have an unique identifier

Two different objects within ***DBLP*** dataset were identified: (i) ***publication*** and (ii) ***cited publication*** to which the reference can be found in at least one of the

publication object. The number of objects in the *DBLP* dataset is shown in Tab. 1.

Object type	No. of objects
cited publication	15,014
publication	1,980,666

Tab. 1. Objects quantity in the experimental dataset

The large difference between number of publications and cited publications may be related to the fact that not all cited publications are indexed by DBLP.

For each object type, only one activity type was detected. Objects with the activities that can be performed by user toward them are shown in Tab. 2. In general, there was one type of activity that user can perform toward an object: activity for object creation (publication authoring). However, we can distinguish two types of authors: (i) these who are authors of publications and (ii) these who are the authors of cited publications i.e. references that were included in the publication objects. The former perform activity called publication authoring and the latter perform cited publication authoring activity (Tab. 2)

Object type	Activity type	No. of activities
cited publication	cited publication authoring	34,910
publication	publication authoring	3,892,785

Tab. 2. Activities assigned to object types

As we can see, there are 1,980,666 publications and 3,892,785 publication authoring activities, that means there are on average 2 authors for 1 publication. Similar situation occurs when the cited publication objects and cited publication authoring activities are considered, Tab. 2.

The next step was to extract from cleansed data the hierarchical pre-social network. The data about users and their activities within HPSN are summarized in Tab. 3.

Activity type	No. of activities	Distinct users with the activity	Total users with a given activity
Publication authoring	3,892,785	793,778	100%
Cited publication authoring	34,910	12,374	1.6%

Tab. 3. Activities in hierarchical pre-social network

Activity type	No. of activities before flattening (in HPSN)	No. of activities after flattening (in FPSN)			
		New	End level: publication (bottom–top approach)	New	End level: cited publication (top–bottom approach)
Publication authoring	3,892,785		38,92,785	+	171,791
Cited publication authoring	34,910	+	169,750		34,910

Tab. 4. Number of activities before and after flattening; '+' denotes activates which were created during a given flattening process

Both top–bottom and bottom–top flattening processes (see Section 4) have been applied to the hierarchical pre-social network HPSN. In the former the end level was publication level and in the latter cited publication level of hierarchy was the end level. After the flattening process some of the activities were multiplied in comparison to HPSN (Tab. 4).

Once the flattening processes have been completed, the three separate layers in each of two created multi-layered social networks were identified and relationships between users on these layers were extracted (see Section 5). Both the layers and the number of distinct relationships between users extracted on each layer are presented in Tab. 5.

Layer	New	No. relationships with common activities on the layer of SN				
		Moved	Publication (bottom–top approach)	Moved	Cited publication (top–bottom approach)	Cited publication / Publication
publication author – publication author			7,836,852	+	12,928,578	1.65
publication author – cited publication author	+		410,818		410,818	1.00
cited publication author – cited publication author		+	6,801,764		78,628	0.01

Sum:	15049434	13418024	2.66
	New: 3%	New: 3%	
	Moved: 45%	Moved: 96%	

Tab. 5. Layers in SN and their profile for two different flattening processes

One new layer *named publication author – cited publication author* in the multi-layered social network was created during the flattening process. This means that new user relationship was detected. Percentage of new relationships in SN is 3% for both publications and cited publications, Tab. 5, Fig. 14. It is worth mentioning that the new relationship would not be visible without the flattening process. Therefore, the method of social network pre-processing, which utilises the concept of flattening for hierarchical object relations, enables to discover new knowledge about the connections between people.

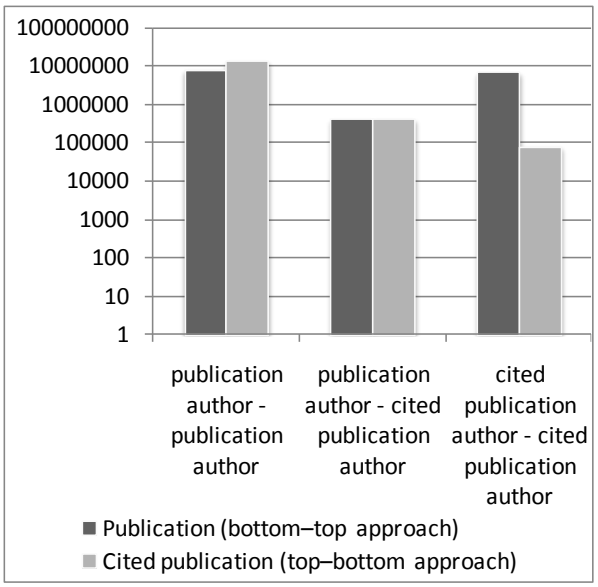

Fig. 14. No. of activities within created layers for three different flattening processes

8. Conclusions and Future Work

The described in the paper approach enables to cope in the systematic way with the data about objects, the hierarchies between them and the activities performed by users towards these objects. It is very important in the context of large-scale and multi-layered social networks analysis.

Future work will focus on analysis of different data sets in the context of their pre-processing for multi-layered social networks especially data about user activities in web-based systems. Another direction of research are studies on efficiency of computing for large data sets, what is very important in some domains [3].

Acknowledgments

The authors are indebted to Elżbieta Kukla for her valuable discussion and Tomasz Filipowski for the dataset preparation. The work was supported by The Polish Ministry of Science and Higher Education, the development project, 2009-2011.

9. References

[1] Agarwal, N., Galan, M., Liu, H., Subramanya S.: WisColl: Collective Wisdom based Blog Clustering, Information Sciences, Vol. 180, Issue 1, 2010, pp. 39-61.

[2] Blondel V.D., Guillaume J.-L., Lambiotte R., Lefebvre E.: Fast unfolding of communities in large networks, J. Stat. Mech., P10008, 2008.

[3] Bródka P., Musiał K., Kazienko P.: A Performance of Centrality Calculation in Social Networks. International Conference on Computational Aspects of Social Networks, CASoN 2009, June 24-27, 2009, Fontainebleau, France, IEEE Computer Society, 2009, pp. 24-31.

[4] Capocci A., Servedio V., Colaiori, F., Buriol L., Donato D., Leonardi S., Caldarelli G.: Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia, Physical Review E, vol. 74, Issue 3, id. 036116, 2006.

[5] Cheng X., Dale C., Liu J.: Statistics and social networking of YouTube videos. 16th International Workshop on Quality of Service, IWQoS 2008, IEEE, 2008, pp. 229-238.

[6] Ellison, N.B., Steinfield, C., Lampe, C.: The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. Journal of Computer-Mediated Communication, 12(4), article 1. http://jcmc.indiana.edu/vol12/issue4/ellison.html, 2007.

[7] Girvan M., Newman M.E.J., Community structure in social and biological networks, Proc. Natl. Acad. Sci., USA, 99 (12) (2002), pp. 7821–7826.

[8] Huberman B., Romero D., Wu F.: Social networks that matter: Twitter under the microscope. First Monday, 2009, pp 1-5 (arXiv:0812.1045v1).

[9] Kazienko P., Bródka P., Musial K.: Individual Neighbourhood Exploration in Complex Multi-layered Social Network, IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology, 31 August - 3 September 2010, Toronto, Canada, submitted.

[10] Kazienko P., Musial K., Kajdanowicz T.: Multidimensional Social Network and Its Application to the Social Recommender System. IEEE Transactions on

Systems, Man and Cybernetics - Part A: Systems and Humans, 2010, in press.
[11] Kazienko P., Ruta D., Bródka P.: The Impact of Customer Churn on Social Value Dynamics. International Journal of Virtual Communities and Social Networking, 1(3), July-September 2009, pp. 60-72.
[12] Lomi A., Lusher D., Pattison P. E., Robins G.: Inter-organizational hierarchies, social networks, and identities in multi-unit organizations, American Sociological Association, TBA, New York, 2007, http://www.allacademic.com/meta/p183413_index.html.
[13] Mika P.: Flink: Semantic Web technology for the extraction and analysis of social networks, Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 3, Issues 2-3, Selected Papers from the International Semantic Web Conference, 2004 - ISWC, 2004, October 2005, pp. 211-223, ISSN 1570-8268.
[14] Matsuo Y., Mori J., Hamasaki M., Nishimura T., Takeda H., Hasida K., Ishizuka M.: POLYPHONET: An advanced social network extraction system from the Web, Web Semantics: Science, Services and Agents on the World Wide Web, Volume 5, Issue 4, World Wide Web Conference 2006 Semantic Web Track, December 2007, pp. 262-278, ISSN 1570-8268.
[15] Scott J.: Social Network Analysis: A Handbook, SAGE Publications, London, UK, 2000.
[16] Tyler J.R., Wilkinson D.M., Huberman B.A.: Email as spectroscopy: Automated discovery of community structure within organizations, in: Communities and Technologies, Kluwer, B.V., Deventer, The Netherlands, 2003, pp. 81-96.
[17] Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press, New York, 1994.