ИССЛЕДОВАНИЕ СТРУКТУРЫ ГРАФА НАУЧНОГО СОАВТОРСТВА МЕТОДАМИ АНАЛИЗА СОЦИАЛЬНЫХ СЕТЕЙ

Басараб М.А.[1], Глинская Е.В.[2], Иванов И.П.[3], Колесников А.В.[4], Кузовлев В.И.[5]

Методы анализа социальных сетей (SNA) являются эффективным инструментарием для качественного и количественного анализа связей в социальном графе, выделения подгрупп и ключевых элементов различного рода, прогнозирования состояний. В работе представлен обзор основных метрик SNA, позволяющих провести базовый анализ экспериментальных данных, выделить ключевые узлы, сделать качественный вывод о структуре сети. В результате оказывается возможным установить значимость отдельных лиц в общей картине социального взаимодействия, выделить подгруппы связанных узлов и принципы распространения информации в сети. В прикладной задаче исследования структуры социальной сети научного соавторства метрики SNA могут помочь в выделении наиболее продуктивных авторов, например, для дальнейшей мотивации, а также прогноза динамики такой сети для задач мониторинга и управления. В качестве примера представлены результаты анализа социальной сети, узлы которой соответствуют сотрудникам одной из кафедр МГТУ им. Н.Э. Баумана, находящимся в соавторстве при публикации статей. В пакете Gephi проведен анализ данных, накопленных с 2012 по 2015 гг. В отличие от других работ по данной тематике, сеть рассматривается как динамическая система, и проводится ее анализ за несколько последующих лет.

***Ключевые слова**: технологии обработки больших массивов, социальный граф (SNA), граф сотрудничества, сеть соавторства, метрики SNA, электронная библиотека, киберпространство, продуктивные авторы.*

DOI: 10.21681/2311-3456-2017-1-31-36

Введение

Исследование технологий обработки больших массивов данных (концепция «извлечения данных» – англ. «data mining») является в настоящее время одним из важнейших направлений в обеспечении широкого спектра задач информационной безопасности, в частности, проблем косвенной доступности информации путем ее получения специализированными средствами анализа «больших данных» (англ. «big data»). К структурам, содержащим информацию такого рода относятся социальные сети.

В методах анализа социальных сетей (англ. SNA – Social Network Analysis) сеть рассматривается в терминах теории графов как система, состоящая из узлов (люди, организации и т.д.) и различного рода связей между ними. Одной из мер сложности таких систем является количество связей между их элементами [1], однако метрики SNA позволяют получить не только статистические данные о связях в социальной сети, но и выделить сильно- и слабосвязанные элементы, определить ключевые узлы сети, выявить подгруппы, распознать паттерны в строении и эволюции сети. Методы SNA, возникнув на стыке социологии, статистики, теории графов, теории сетей и обработки больших массивов данных, нашли широкое применение во множестве сфер общественных отношений, от политологии, до борьбы с терроризмом, особенно на фоне бурного развития всевозможных онлайновых социальных сетей [2-4].

Одним из известных примеров социальных сетей является сеть соавторства или граф сотрудничества (англ. - Collaboration graph) [5, 12, 13], узлами которого являются участники сети, а ребро между двумя узлами говорит о наличии некоторого отношения ними. Одним из известнейших примеров таких графов является граф соавторства математиков или граф Эрдёша [5, 6].

В работе представлен обзор основных метрик SNA, позволяющих провести базовый анализ экспериментальных данных, выделить ключевые узлы, сделать качественный вывод о структуре сети. Представлены результаты анализа социальной сети, узлы которой соответствуют сотрудникам одной из кафедр МГТУ им. Н.Э. Баумана, на-

1 Басараб Михаил Алексеевич, доктор физико-математических наук, профессор, МГТУ им. Н.Э. Баумана, Москва, bmic@mail.ru
2 Глинская Елена Вячеславовна, МГТУ им. Н.Э. Баумана, Москва, glinskaya-iu8@rambler.ru
3 Иванов Игорь Потапович, доктор технических наук, МГТУ им. Н.Э. Баумана, Москва, ivanov@bmstu.ru
4 Колесников Александр Владимирович, кандидат технических наук, МГТУ им. Н.Э. Баумана, Москва, avkolesnikov90@list.ru
5 Кузовлев Вячеслав Иванович, кандидат технических наук, доцент, МГТУ им. Н.Э. Баумана, Москва, v.kuzovlev@bmstu.net

ходящимся в соавторстве при публикации статей. На основе информации, опубликованной на сайте электронной библиотеки МГТУ им. Н.Э.Баумана[6], в пакете Gephi [7] проведен анализ данных, накопленных с 2012 по 2015 гг. Исследование динамического поведения сети отличает данную работу от ряда других исследований подобной тематики (см., например, [8]).

Обзор основных метрик SNA

SNA располагает рядом метрик для определения ключевых узлов, общей значимости отдельных узлов, сильносвязанных узлов, групп и подгрупп, а также направлений распространения информации в сети. Перечислим далее основные метрики SNA [2].

Степень центральности (англ. *degree centrality*)

$$D_i = \frac{k_i}{N-1} = \frac{\sum a_{ij}}{N-1}$$

определяется по матрице связности $[a_{ij}]$ графа, соответствующего социальной сети *G*, и равна количеству смежных вершин, нормированных на максимально возможное их количество $(N-1)$; не зависит от размера сети N и изменяется в интервале от 0 до 1; согласно гипотезе вершина с большим значением параметра D_i имеет высокую степень активности и информационного влияния в своей окрестности.

Степень промежуточности (англ. *betweenness centrality*)

$$B_i = \frac{\sum_{i<k\in G} n_{jk}(i)/n_{jk}}{(N-1)(N-2)},$$

где n_{jk} - количество всех кратчайших путей между узлами *j* и *k*; $n_{jk}(i)$ - количество кратчайших путей между узлами *j* и *k*, проходящих через узел *i*.

Узлы с высоким значением показателя B_i (так называемые «информационные брокеры») играют ключевую роль в распространении информации между двумя или более плотно связанными подмножествами сети; удаление таких вершин может оказать сильное негативное воздействие на всю сеть.

Степень близости (англ. *closeness centrality*)

$$C_i = (L_i)^{-1} = \frac{N-1}{\sum_{j\in G} d_{ij}},$$

где d_{ij} - расстояние между узлами *i* и *j*; L_i - нормированное расстояние между узлом *i* и другими вершинами графа. Через узлы с малой степенью близости можно легче и быстрее достичь остальных вершин графа.

Метрика собственных векторов (англ. *eigenvector centrality*)

$$x_i = \frac{1}{\lambda}\sum_{j\in M(i)} x_j = \frac{1}{\lambda}\sum_{j\in G} a_{ij}x_j,$$

где $M(i)$ - набор соседей вершины *i*; λ - константа (собственное число). Данная метрика определяет, насколько хорошо узел *i* связан с другими хорошо связанными узлами.

Индекс PageRank [3] дает возможность сравнить относительную «важность» узлов на основе аналогии между гиперссылками веб-страниц и связями в сети. Рекурсивно определяется как

$$PR(A) = (1-d) + d\sum_{B\in M(A)} \frac{PA(B)}{L(B)},$$

где $M(A)$ – множество соседей узла *A*; $L(B)$ – количество исходящих из узла B связей; d – коэффициент затухания.

Обзор других типов метрик и различных типов исследуемых графов приведен в [4]. В работе [9] приведен достаточно полный обзор программных средств для SNA и визуализации сетевых графов.

Для визуализации социального графа наиболее популярными в настоящее время являются алгоритмы, основанные на аналогиях физических принципов притяжения и отталкивания тел или частиц по закону Гука, Кулона и др. с целью минимизации энергии системы (Force-directed graph drawing) [10]. В частности, одним из широко используемых алгоритмов является Force Atlas 2 [11] и специализированное ПО Gephi [7], находящееся в открытом доступе.

Сбор и предварительный анализ данных

Сбор данных по публикациям проводился в течение 4 лет на одной из кафедр МГТУ им. Н.Э. Баумана по данным электронной библиотеки. Общие статистические данные по годам представлены в (табл.1), где узел – соавтор публикации, а связь между узлами формируется, если установлено соавторство ученых хотя бы в одной работе.

На основе собранных данных мы можем сделать весьма поверхностный вывод о ежегодном увеличении числа публикуемых авторов при сохранении количества публикаций.

Выделить ключевые узлы в сети, а также сделать некоторое количественное заключение об изменении характера сети научного соавторства помогут метрики SNA.

6 Библиотека МГТУ им. Н.Э. Баумана. URL: http://library.bmstu.ru

Таблица 1.
Общая статистика узлов и связей сети соавторства

	2012	2013	2014	2015	Всего
Узлы	27	28	29	30	41
Ребра	54	53	43	56	150

Анализ данных методами SNA

В первую очередь визуализируем графы, соответствующие экспериментальным данным. Используем для этого программный пакет Gephi (рис.1).

Данные предварительно были обезличены. Аббревиатуры на (рис.1) соответствуют

«проф» – профессор;
«доц» – доцент;
«стпр» – старший преподаватель;
«асс» – ассистент.

Номер «№» указывает на порядковый номер сотрудника на данной должности по алфавиту в штатном расписании.

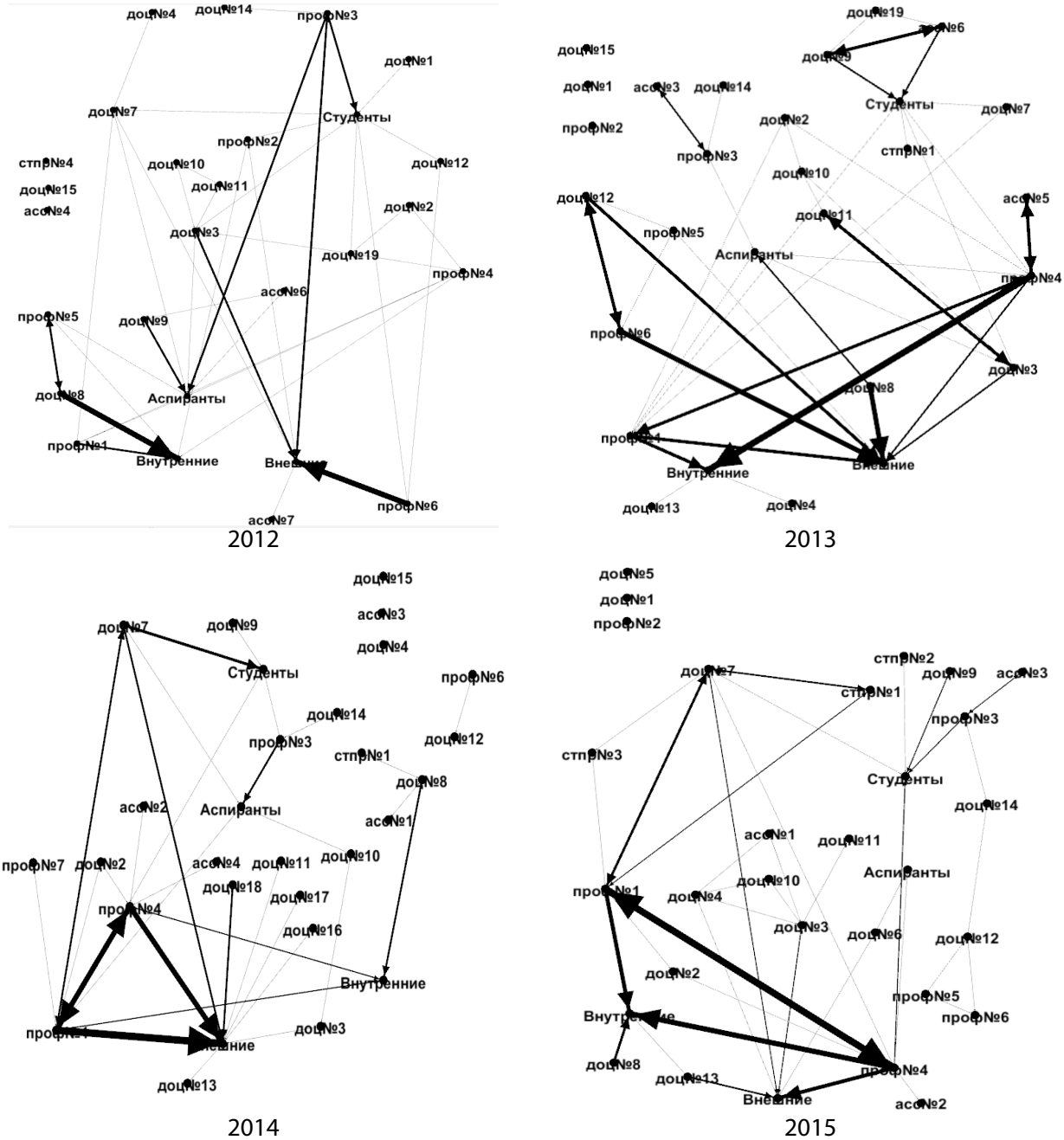

***Рис.1**. Графы, соответствующие сети научного соавторства, по годам*

Таблица 2.
Основные метрики SNA, по годам

Год	Степень центральности	Взвешенная степень центральности	Плотность графа	Глобальный коэффициент кластеризации	Средняя степень промежуточности
2012	4,296	2,741	0,083	0,258	2,175
2013	4,429	3,929	0,082	0,328	1,630
2014	3,724	3,552	0,067	0,218	1,707
2015	4,133	4,267	0,071	0,315	1,681
Общее	6,927	4,756	0,087	0,340	3,054

Только на основе визуальной оценки графов уже можно сделать несколько достаточно интересных выводов, если учесть, что толщина связи на рисунке учитывает вес ребра графа, то есть количество работ в течение года, в которых были указаны одни и те же соавторы. Большинство публикаций связаны соавторством с «внутренними» сотрудниками (учеными из того же университета), внешними соавторами (сотрудники других организаций) и студентами. По весам, то есть количеству публикаций, выделяются доц№8, проф№4, проф№1.

В заключении визуального анализа графов отметим, что из года в год остается 3 публикации на кафедре, выполненные без соавторов, разными учеными, за одним исключением – доц№15 отмечен на трех графах из четырех.

Перейдем к анализу основных метрик SNA (табл.2).

Усредненная степень центральности намеренно не была нормирована, т.к. оценивались показатели по разным годам, обладающим узлами с разным максимальным числом связей.

Рейтинг узлов по некоторым метрикам представлен в (табл.3).

На основании полученных данных можно сделать ряд выводов. Во-первых, в 2014 г. мы видим явный минимум количества публикаций – было снижено количество связей между узлами, что отражает показатель степени центральности; получены минимальные значения плотности графа и среднего коэффициента кластеризации; при этом локально вырос показатель средней степени промежуточности, который в совокупности с оценкой рейтинга взвешенной мощности узлов (количества входящих и исходящих связей) позволяет сделать вывод о снижении числа публикаций для всей сети, кроме проф. №1 и проф№4, степень научного соавторства которых возросла и увеличила показатель средней промежуточности. Во-вторых, плотность графа снижалась на протяжении всей выборки, число исходящих связей с «внешними» соавторами увеличивалось, а число связей с «внутренними» (сотрудниками того же университета) сохранялась на прежнем уровне, что говорит об увеличении числа контактов публикуемых авторов с другими университетами. Коэффициент кластеризации в анализе используется чаще плотности графа, однако в данном случае небольшие размеры выборки и самой сети не позволяют сделать какой-то однозначный вывод о кластеризации узлов, можно лишь заметить, что количественно показатель принимает небольшие значения, что говорит об относительно равномерном распределении связей в графе.

Показатели некоторых метрик для отдельных узлов позволяют выделить лидеров сети по количеству исходящих связей – то есть насколько часто

Таблица. 3
Рейтинг узлов по различным метрикам

Исходящие связи		Степень промежуточности		Метрика собственных векторов	
1	проф№4	1	доц№8	1	Внешние
2	проф№1	2	проф№4	2	Студенты
3	доц№3	3	проф№5	3	Аспиранты
4	доц№7	4	доц№3	4	проф№4
5	доц№8	5	доц№19	5	проф№1

авторы статьи указывали других ученых в качестве соавторов – это проф№4, проф№1 и доц№3. Согласно степени промежуточности, отметим значимость доц№8, проф№4, проф№5 – это те узлы, через которые проходит максимальное число кратчайших путей между всеми узлами сети.

Заключение

Методы SNA предоставляют уникальный набор инструментов для обработки сложно структурированных больших объемов данных, позволяют не только качественно, но и количественно оценить характер протекающих процессов, заглянуть за рамки привычной статистики. Работа с социальной сетью с точки зрения теории графов позволяет установить значимость отдельных лиц в общей картине социального взаимодействия, выделить подгруппы связанных узлов и принципы распространения информации в сети. В прикладной задаче исследования структуры социальной сети научного соавторства метрики SNA могут помочь в выделении наиболее продуктивных авторов, например, для дальнейшей мотивации, а также прогноза динамики такой сети для задач мониторинга и управления. В дальнейшем интересной видится задача исследования такой социальной сети в динамике для всей организации, что потребует обработки гораздо больших объемов данных.

***Рецензент**: Матвеев Валерий Александрович, доктор технических наук, профессор, заведующий кафедры «Информационная безопасность» МГТУ им. Н.Э. Баумана, va.matveev@bmstu.ru*

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-29-09517 офи_м.

Литература

1. Анфилатов В.С., Емельянов А.А., Кукушкин А.А. Системный анализ в управлении: Учебное пособие / Под ред. А.А. Емельянова. М.: Финансы и статистика, 2002. 368 с.
2. Freeman L.C. Centrality in social networks conceptual clarification // Social networks. – 1979. – 1(3). – P. 215-239.
3. Page L., Brin S., Motwani R., Winograd T. The pagerank citation ranking: bringing order to the web // January 29, 1998. URL: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf (Дата обращения 01.07.2016).
4. Hopkins A. Graph theory, social networks and counter terrorism. May 19, 2010. Univ. of Massachusetts Dartmouth. URL: https://compmath.files.wordpress.com/2010/05/ ahopkins_freports10.pdf (Дата обращения 01.07.2016).
5. Yegnanarayanan V., Umamaheswari G.K. A Note on the Importance of Collaboration Graphs // Int. J. of Mathematical Sciences and Applications // September 2011. - Vol. 1. - No. 3. – P.1113-1121.
6. Batagelj V., Mrvar A. Some analyses of Erdos collaboration graph // Social Networks. - 2000. - Vol. 22. - No.2. - P. 173–186.
7. Bastian M., Heymann S. Gephi: an open source software for exploring and manipulating networks, ICWSM. – 8. – P. 361-362.
8. Apanovich Z.V., Vinokurov P. S., Elagin V.A. An approach to visualization of knowledge portal content // Bulletin of NCC. – Issue 29. – 2009. – P.17-32.
9. Choudhary P., Singh U. A survey on social network analysis for counter-terrorism // Int. Journal of Computer Applications. – Feb. 2015. – Vol. 112. – No. 9. – P. 24-29.
10. Kobourov, S.G. Spring Embedders and Force-Directed Graph Drawing Algorithms. 2012. URL: http://arxiv.org/abs/1201.3011 (Дата обращения 01.07.2016).
11. Jacomy M., Heymann S., Venturini T., Bastian M. ForceAtlas2, a graph layout algorithm for handy network visualization // Paris. August 29, 2011. URL: http://webatlas.fr/ tempshare/ForceAtlas2_Paper.pdf (Дата обращения 01.07.2016).
12. Fernandes J.M., Authorship trends in software engineering // Scientometrics. – 2014. – 101. – P. 257-271. doi:10.1007/s11192-014-1331-6.
13. Arif. T., Mining and Analyzing Academic Social Networks // Int. J. of Computer Applications Technology and Research. – 2015. – Vol. 4. – Issue 12. – P. 878-883.

STUDY INTO THE STRUCTURE OF THE SCIENTIFIC COATHORSHIP GRAPH USING SOCIAL NETWORK ANALYSIS

Basarab M.[7], Glinskaya E.[8], Ivanov I.[9], Kolesnikov A.[10], Kuzovlev V.[11]

Abstract.*Social network analysis (SNA) is an effective tool for qualitative and quantitative analysis of connections in the social graph, identifying subgroups and various key elements, and forecasting conditions. This article reviews the main SNA metrics that enable basic analysis of the experimental data, identifying key nodes, and making qualitative conclusion about the network structure. This makes it possible to establish significance of certain people in the general picture of social interaction, identify subgroups of companion nodes and principles of the information distribution in the network. The application task of the study into the structure of the social network of the scientific co–authorship of SNA metrics can help to identify the most productive authors, for instance, for further motivation, and also forecasting dynamics of such network for monitoring and management tasks. For instance, we provide results of analysis into social network, which nodes correspond to the employees, co–authors of the published articles, working at one of the departments in N.E. Bauman Moscow State Technical University. Gephi package analyses data accumulated from 2012 to 2015. Unlike other works on this topic, the network is viewed as a dynamic system, and its analysis covers several years.*

Keywords: *secure communication, social network, social network analysis, social graph, collaboration graph, co–authorship graph, co–authorship network analysis metrics, social interaction, cyberspace, dy–namic system, electronic library*

References

1. Anfilatov V.S., Emel'yanov A.A., Kukushkin A.A. Sistemnyy analiz v upravlenii: Uchebnoe posobie, by ed. A.A. Emel'yanov. Moscow, Finansy i statistika, 2002. 368 P.
2. Freeman L.C. Centrality in social networks conceptual clarification, Social networks, 1979, 1(3), pp. 215-239.
3. Page L., Brin S., Motwani R., Winograd T. The pagerank citation ranking: bringing order to the web, January 29, 1998. URL: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.
4. Hopkins A. Graph theory, social networks and counter terrorism. May 19, 2010. Univ. of Massachusetts Dartmouth. URL: https://compmath.files.wordpress.com/2010/05/ ahopkins_freports10.pdf.
5. Yegnanarayanan V., Umamaheswari G.K. A Note on the Importance of Collaboration Graphs, Int. J. of Mathematical Sciences and Applications, September 2011, Vol. 1, No. 3, pp. 1113-1121.
6. Batagelj V., Mrvar A. Some analyses of Erdos collaboration graph, Social Networks, 2000, Vol. 22, No.2, pp. 173–186.
7. Bastian M., Heymann S. Gephi: an open source software for exploring and manipulating networks, ICWSM, 8, P. 361-362.
8. Apanovich Z.V., Vinokurov P. S., Elagin V.A. An approach to visualization of knowledge portal content, Bulletin of NCC, Issue 29, 2009, P.17-32.
9. Choudhary P., Singh U. A survey on social network analysis for counter-terrorism. Int. Journal of Computer Applications, Feb. 2015, Vol. 112, No. 9, P. 24-29.
10. Kobourov, S.G. Spring Embedders and Force-Directed Graph Drawing Algorithms. 2012. URL: http://arxiv.org/abs/1201.3011.
11. Jacomy M., Heymann S., Venturini T., Bastian M. ForceAtlas2, a graph layout algorithm for handy network visualization, Paris. August 29, 2011. URL: http://webatlas.fr/ tempshare/ForceAtlas2_Paper.pdf.
12. Fernandes J.M., Authorship trends in software engineering. Scientometrics, 2014, 101, P. 257-271. doi:10.1007/s11192-014-1331-6.
13. Arif. T., Mining and Analyzing Academic Social Networks. Int. J. of Computer Applications Technology and Research, 2015, Vol. 4, Issue 12, pp. 878-883.

7 Basarab Mikhail, Dr.Sc., Professor, Bauman Moscow State Technical University, Moscow, bmic@mail.ru
8 Glinskaya Elena, Bauman Moscow State Technical University, Moscow, glinskaya-iu8@rambler.ru
9 Ivanov Igor, Dr.Sc., Bauman Moscow State Technical University, Moscow, ivanov@bmstu.ru
10 Kolesnikov Aleksandr, Ph.D., Bauman Moscow State Technical University, Moscow, vkolesnikov90@list.ru
11 Kuzovlev Vyacheslav, Ph.D., Assistant Professor, Bauman Moscow State Technical University, Moscow, v.kuzovlev@bmstu.net