Анализ методов построения графа соавторства: подход на основе двудольного графа

Ф.В.Краснов

***Аннотация*— Сложившаяся практика построения графов соавторства подразумевает использование математического аппарата теории графов. Традиционно для построения графов соавторства используют неориентированные графы. Авторами данного исследования проведен анализ использования двудольного ориентированного графа в качестве инструмента для построения графов соавторства. В исследовании показаны преимущества от использования двудольного графа и проведено количественное сравнение традиционного способа построения графа соавторств и способа на основе двудольного графа с использованием метрик центральности графов.**

***Ключевые слова*— двудольный граф, граф соавторств, метрики центральности графа, визуализация графа, RANSAC.**

I. Введение

Одним из первых исследований графа соавторства является работа [1], сделанная в 1974 году. С этого времени исследования научной деятельности при помощи графов соавторства не прекращались и обрели статус проверенного инструмента анализа. Например, в недавнем исследовании [2] предпринята попытка предсказания будущих соавторов на основе графа соавторства, а в работе [3] построен глобальный граф соавторства на основе Google Scholar, который содержит более 400 тысяч вершин. Оба исследования [2,3] проведены в 2017 году.

Построение графа соавторства выполняется таким образом, что если два автора сделали совместную научно-исследовательскую работу, то каждый из авторов считается вершиной графа, а факт соавторства – ребром. Будем называть такой способ создания графа соавторств традиционным. В результате такого, общепринятого подхода, например, автора исследования [4] получают граф, изображенный на рисунке (Рисунок 1).

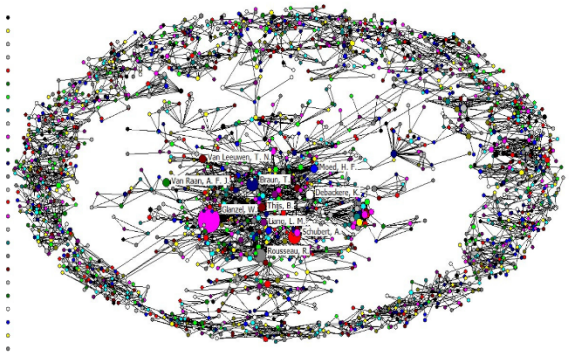

Рисунок 1. Граф соавторства [4].

Полученный на (Рисунок 1) граф представляет наглядную визуализацию выбранного научного сообщества и позволяет производить анализ с помощью таких распространенных метрик графов, как: *Betweenness centrality* [5, 8, 10] и *Closeness centrality* [6,7,9]. Данные метрики, как и метрика *Degree*, предназначены для формального выделения важных вершин графа.

Существенным аспектом для построения графа соавторств является выборка данных для анализа. Обычно исследователи используют публичную библиографическую информацию, содержащую список соавторов. Источником такой информации может быть Google Scholar [3], ArXiv [11] и другие онлайн библиотеки [12, 13]. Рассмотрение открытых научных сообществ так же интересно, как и сужение выборки до одной страны [15], отрасли [14] и даже организации [16].

Добавление в граф полей, связанных с аффиляцией автора, позволяет сделать исследования отношения организаций. Как пример, в работе [14] авторы анализируют связи между исследовательскими институтами и индустриальными научными центрами в нефтегазовой индустрии. Такой подход к выборке позволяет проанализировать топологию связей между организациями, на основании принадлежности авторов к организации (Рисунок 2).

Статья получена 14 декабря 2017.

Ф.В.Краснов, к.т.н., эксперт, ООО «Газпромнефть НТЦ», г. Санкт-Петербург, набережная реки Мойки д.75-79, 190000. krasnov.fv@gazprom-neft.ru, orcid.org/0000-0002-9881-7371, РИНЦ 8650-1127

Рисунок 2. Граф исследовательских организаций [14].

Отметим, что все приведенные выше исследования не принимают в расчет содержание исследовательских статей. Эта особенность будет важна в дальнейшем.

Среднее количество соавторов может изменяться в зависимости от индустрии [17], но в целом количество соавторов растет. Отметим этот факт, как структурную особенность исследуемой области.

В приведенных выше исследования граф соавторства строится на неориентированный граф. Авторы равнозначны в соавторстве, хотя на деле это не так. В работе [18] проанализирована структура команды соавторов и сформулированы возможные роли в процессе исследования.

Кроме того, в традиционном построении графа соавторства информация о всех совместных исследовательских работа содержится в ребрах графа. Часто ребра рисуют различной толщины или цвета (Рисунок 2) в зависимости от количества совместных работ, но данная характеристика ребер не рассматривается в контексте метрик графа, так как не отражает коммуникационный смысл повторного соавторства. С учетом этих ограничений сформулируем следующий исследовательские вопросы:

И.В.1: Существуют ли другие способы построения графа соавторств?

И.В.2: Какими преимуществами и недостатками обладают различные способы построения графа соавторств?

И.В.3: Каковы количественные, сравнимые характеристики графов соавторств?

В данном исследовании предпринята попытка ответить на поставленные исследовательские вопросы. Исследование состоит из введения, описания методики, предложенной авторами, анализа от применения данной методики к результатам научной работы отраслевого научно-технического центра компании «Газпромнефть» и полученным выводам.

II. Методика исследования

В приведенных выше исследованиях граф соавторства строится как неориентированный граф: статьи становятся равнозначными ребрами, соединяющими авторов. Авторы данного исследования считают, что более информативным будет построение графа соавторств как двудольного графа. Такой подход позволяет включить в граф соавторств информацию о научных статьях. На рисунке (Рисунок 3) приведен основной принцип построения графа соавторств на основе направленного двудольного графа.

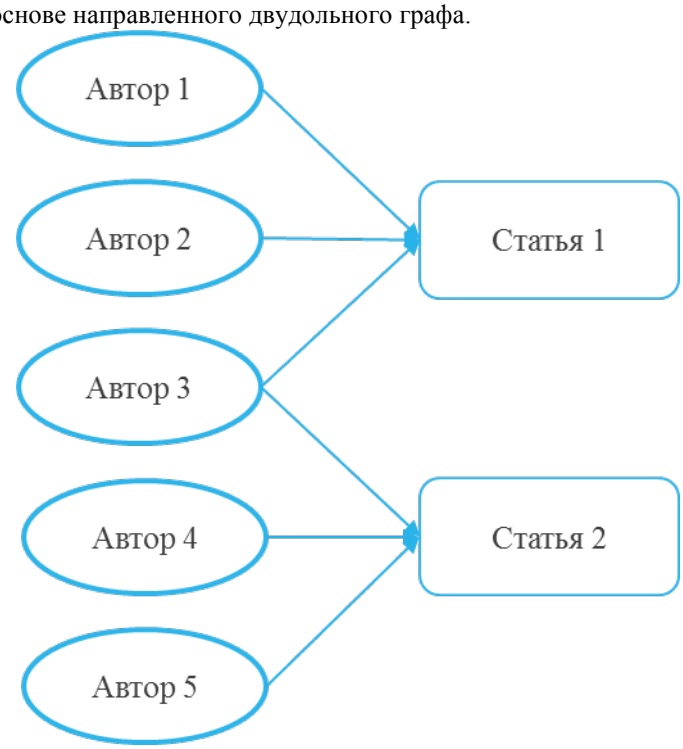

Рисунок 3 Двудольный граф соавторств.

Преимущества такого подхода состоят в том, что в графе соавторств становится возможным сохранить для дальнейшего анализа библиографическую информацию о статье:

1. Название статьи
2. Год издания
3. Издатель
4. Ключевые слова

Отметим, что традиционное представление графа соавторств в виде неориентированного графа является проекцией двудольного графа на множество вершин авторов. Поясним это более подробно.

Ориентированный граф $G = (V, E)$ называется двудольным, если множество его вершин можно разбить на две части $A \cup P = V$, так, что

- ни одна вершина в A не соединена с вершинами в P и
- ни одна вершина в P не соединена с вершинами в A.

В нашем случае A- это множество авторов, а P- это множество статей. A и P- являются долями графа G. Отметим, что граф G может быть как полным так и неполным в зависимости от того имеют ли авторы соединения со всеми статьями. Приведенный на (Рисунок 3) двудольный граф является неполным.

Обозначим G_A проекцию графа G на множество вершин A. Граф G_A является традиционным представлением графа соавторств и отображен на рисунке (Рисунок 4).

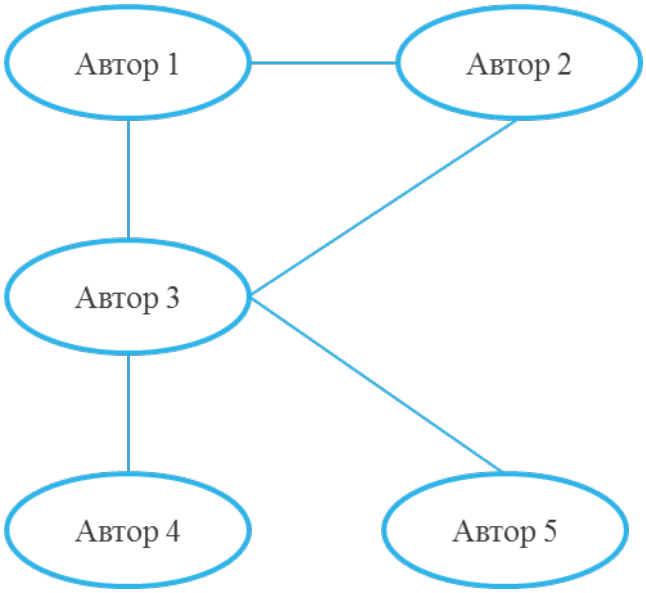

Рисунок 4. Неориентированный граф соавторств.

При построении проекции атрибутами ребер графа G_A могут стать только интегральные характеристики соавторства, например, количество соавторств двух авторов.

III. Сравнение способов построения графа соавторств

Для анализа предложенной авторами методики построения графа соавторств были выбраны статьи, опубликованные сотрудниками научно-технического центра «Газпромнефть» в электронной библиотеке *Onpetro* сообщества нефтегазовых инженеров SPE.

Выборка содержит 140 статей, написанных 385 авторами за 5лет. Среди авторов есть как сотрудники «Газпромнефть НТЦ», так и сотрудники организаций партнеров, с которыми проводились совместные исследования.

Проведем сравнение основных характеристик графа соавторства для традиционного и двудольного построения.

Основная цель анализа научных сообществ с помощью инструментария графа соавторств состоит в том, чтобы определить основные количественные характеристики сообщества. Наиболее важной характеристикой сообщества являются определяющие сообщество индивидуумы. В графе соавторств такие индивидуумы являются важными вершинами. Для поиска таких важных вершин в теории графов определены метрики центральности графов.

Одной из наиболее показательных и распространенных метрик центральности графов является *Betweenness centrality* (C_b) [19]. Эта метрика отражает потенциальные возможности вершины по контролю коммуникаций в графе. $C_b(v)$ определяется как сумма кратчайших путей, которые проходят через данную вершину v, нормированная на общее число кратчайших путей в графе.

C_b описывается следующей формулой:

$$C_b(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

где V полный список вершин графа, $\sigma(s,t)$ количество кратчайших путей от вершины s до вершины t, а $\sigma(s,t|v)$ количество кратчайших путей от вершины s до вершины t, проходящих через вершину v при $v \neq s$ и $v \neq t$.

Следующей по информативности является метрика *Closeness centrality* (C_c) [22], которая является индикатором важности вершины по шкале близости к центру графа. $C_c(v)$ вычисляется как:

$$C_c(v) = \frac{n-1}{V-1} \frac{n-1}{\sum_{u=1}^{n-1} \sigma(u,v)}$$

где $\sigma(v,u)$ количество кратчайших путей от вершины v до вершины u, n – количество вершин, из которых есть возможность достичь вершину v, а V – это количество вершин в графе.

Для вычисления C_bи C_c мы воспользовались алгоритмом из работ [20] и [23] соответственно. Полученные результаты приведены в таблице (Таблица 1).

Таблица 1 Сравнения метрик графа соавторства при разных способах построения.

Метрика/Способ	*Двудольный способ построения графа соавторств*	*Традиционный способ построения графа соавторств*
Betweenness centrality (C_b)	Khasanov, M.M. (0.19) Kovalenko, I.V. (0.17) Zhukov, V.V. (0.17)	Kovalenko, I.V. (0.18) Zhukov, V.V. (0.16) Khasanov, M.M. (0.15)
Closeness centrality (C_c)	OIJ-2016-12-048-051-RU (0.17) Ovcharenko, Yu.V. (0.16) SPE-182031-RU (0.16)	Ovcharenko, Yu.V. (0.32) Belozerov, B.V. (0.32) Zhukov, V.V. (0.32)

Из Таблицы 1 мы видим, что для метрики C_b состав тройки вершин с наибольшими значениями не изменился от способа построения графа соавторств, но наибольшее значение метрики в случае двудольного построения находится в лучшем соответствии с действительностью, так как Khasanov, M.M. является научным лидером данного сообщества согласно [21]. Особенностью метрики C_c для двудольного построения графа соавторства является появления вершин типа статья. Другими словами – это означает, что в центре графа стоит не автор, а сама научная работа.

Сравним визуализацию традиционного и двудольного графа соавторств, представленную на рисунках (Рисунок 5, Рисунок 6).

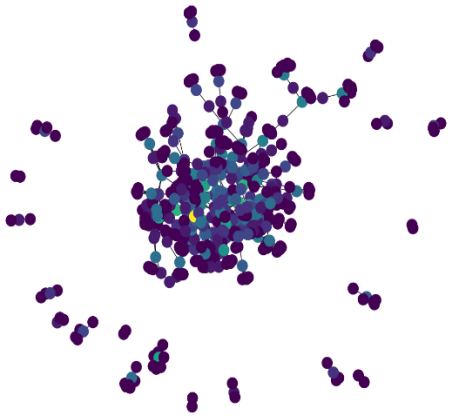

Рисунок 5 Двудольный граф соавторств.

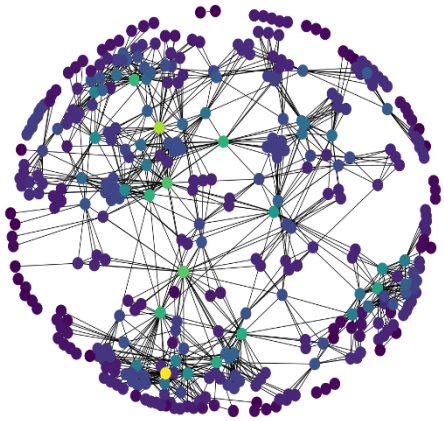

Рисунок 6 Традиционный граф соавторств.

Важно отметить качественную схожесть визуализаций графов соавторств, полученных традиционным способом на (Рисунок 1) и (Рисунок 6).

На (Рисунок 6) и (Рисунок 5) желтым цветом отмечены вершины с наибольшим показателем *Degree*. Из представленной визуализации можно сделать наблюдение о том, что двудольный способ представляет менее информационно перегруженную картину. Так же в визуализации графа, построенного двудольным способом, отчетливо прослеживаются связанные компоненты графа (*Connected components*). В традиционно построенном графе соавторств содержится 15 *Connected components*, а при двудольном построении 20 *Connected components.*

Проанализируем плотность метрики *Degree* для обоих способов построения.

На рисунке (Рисунок 7) отображены плотности метрики *Degree* для обоих способов построения графа соавторства.

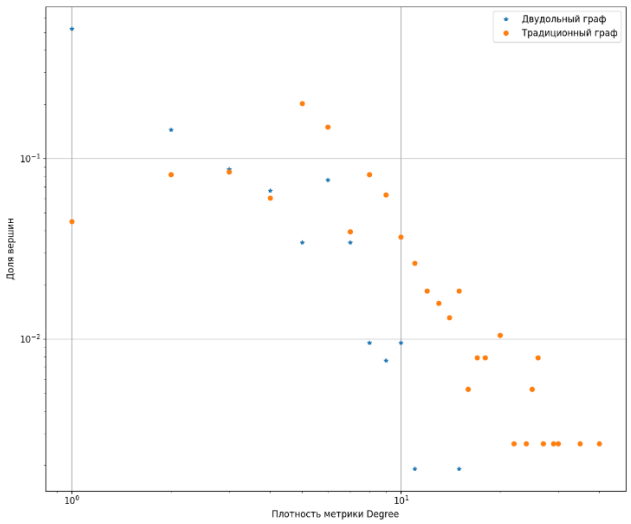

Рисунок 7 Плотность метрики Degree для графов соавторства.

Метрика *Degree* вычисляется как сумма ребер для каждой вершины. На рисунке (Рисунок 7) оси отображены в логарифмическом масштабе. При таком отображении видно, что зависимость носит линейный характер. Сравним качество линейности зависимостей для обоих способов построения графов соавторства. Для этого аппроксимируем зависимости с помощью линейной регрессии.

На рисунках отображена кривая аппроксимации зависимости доли вершин от плотности метрики *Degree* в логарифмических координатах для традиционного (Рисунок 9) и двудольного способа (Рисунок 8) построения графа соавторств.

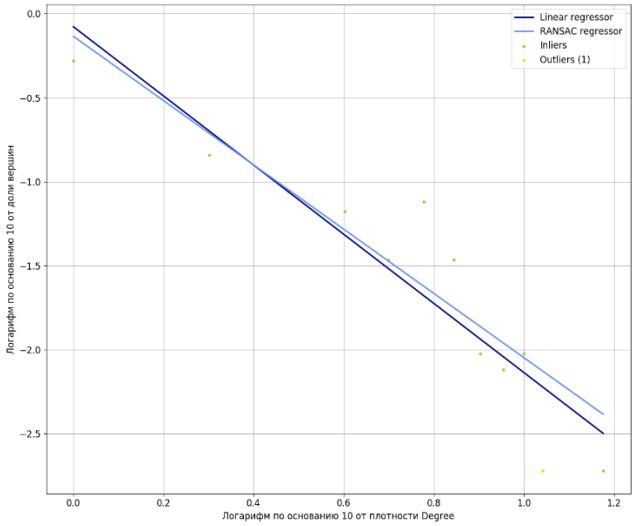

Рисунок 8 Зависимость доли вершин от метрики Degree в логарифмических координатах для двудольного способа построения графа соавторств.

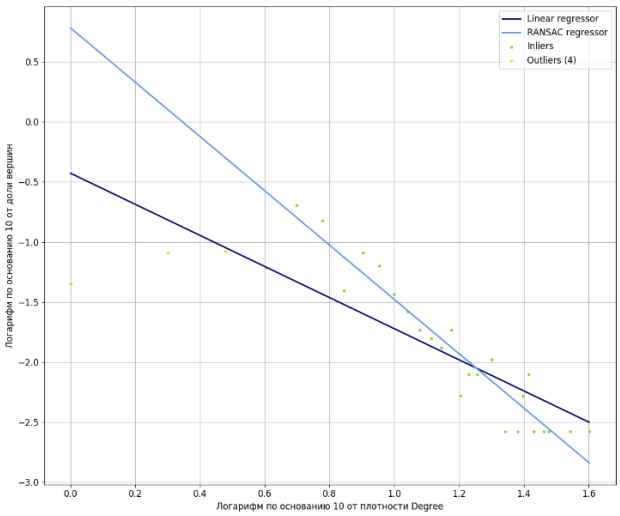

Рисунок 9 Зависимость доли вершин от метрики Degree в логарифмических координатах для традиционного способа построения графа соавторств.

Скоринг для обоих аппроксимаций составил более 86% по метрике R^2, что говорит о достаточно точной линейной аппроксимации зависимостей. Точки, не вписывающиеся в линейную зависимость (outlets), проанализированы с помощью метода RANSAC [24]. Видно, что для традиционного способа построения графа соавторств выбросов (outlets) получается больше.

Наличие линейной зависимости логарифмов двух величин, означает, что между сами величинами присутствует степенная зависимость вида:

$$P(k) = Ck^{-\alpha}$$

Полученная форма зависимость находится в согласии с моделью распространения графов *preferential attachment*, изложенной в исследовании [25]. Таким образом, можно сделать вывод о соответствии полученных обоими способами графов соавторства модели *preferential attachment*, описывающей социологические научные процессы и, в частности, процессы цитирования статей [26].

IV. Заключение

Авторами предложен алгоритм построения графа соавторств на основе двудольного графа. Произведено сравнение двух методов построения графов соавторства: традиционного и на основании двудольного графа.

Сравнение двух способов проведено на основании метрик центральности графов:

- *Betweenness centrality,*
- *Closeness centrality,*
- *Degree*

Сравнение сделанная визуализации графов показывает качественные преимущества способа построения графа соавторств на основании двудольного графа: уменьшение информационной перегрузки, возможность визуального выделения *Connected components*.

Авторами проанализирована зависимость плотности метрики *Degree* и показано, что при обоих способах построения графа соавторства полученные графы соответствуют общим зависимостям социальных процессов, лежащих в их основе.

К существенным преимуществам двудольного способа построения графа соавторства относятся возможность сохранения и анализа информации о научных статьях. В традиционном способе построения графов соавторства такая возможность существует в ограниченном смысле.

В заключении важно отметить, что традиционный способ построения графов соавторства является частным случаем двудольного графа и может быть получен проецированием двудольного графа на плоскость вершин авторов.

Библиография

1. Mullins N. C. The development of specialties in social science: The case of ethnomethodology //Science Studies. – 1973. – Т. 3. – №. 3. – С. 245-273.
2. Chuan P. M. et al. Link prediction in co-authorship networks based on hybrid content similarity metric //Applied Intelligence. – С. 1-17.
3. Chen Y. et al. Building and Analyzing a Global Co-Authorship Network Using Google Scholar Data //Proceedings of the 26th International Conference on World Wide Web Companion. – International World Wide Web Conferences Steering Committee, 2017. – С. 1219-1224.
4. Wei F. et al. A co-authorship network-based method for understanding the evolution of a research area: A case of information systems research //Malaysian Journal of Library & Information Science. – 2017. – Т. 22. – №. 2. – С. 1-14.
5. Leifeld P. et al. Collaboration patterns in the German political science co-authorship network //PloS one. – 2017. – Т. 12. – №. 4. – С. e0174671.
6. Ahmed T. et al. Analysis of co-authorship in computer networks using centrality measures //Communication, Computing and Digital Systems (C-CODE), International Conference on. – IEEE, 2017. – С. 54-57.
7. Chang H. J., Wang W. M. The Hidden Power of Social-Linkage in the Office: A Co-authorship Network Analysis //Proceedings of the 4th Multidisciplinary International Social Networks Conference on ZZZ. – ACM, 2017. – С. 4.
8. Köseoglu M. A. et al. Authorship trends, collaboration patterns, and co-authorship networks in lodging studies (1990–2016) //Journal of Hospitality Marketing & Management. – 2017. – №. just-accepted.
9. Paraschiv I. C. et al. Semantic Similarity versus Co-authorship Networks: A Detailed Comparison //Control Systems and Computer Science (CSCS), 2017 21st International Conference on. – IEEE, 2017. – С. 566-570.
10. Ho T. M. et al. with basic network measures of 2008-2017 Scopus data [version. – 2017.
11. Liu X. et al. Co-authorship networks in the digital library research community //Information processing & management. – 2005. – Т. 41. – №. 6. – С. 1462-1480.
12. Wei F. et al. A co-authorship network-based method for understanding the evolution of a research area: A case of information systems research //Malaysian Journal of Library & Information Science. – 2017. – Т. 22. – №. 2. – С. 1-14.
13. Zhang D. et al. Co-authorship Networks in Additive Manufacturing Studies Based on Social Network Analysis //British Journal of Applied Science & Technology. – 2016. – Т. 15. – №. 1. – С. 1.
14. Gielfi G. G. et al. University-industry research collaboration in the Brazilian oil industry: the case of Petrobras //Rev. Bras. Inov. – 2017. – Т. 16. – №. 2. – С. 325-350.
15. Krasnov F., Yavorskiy R. Measurement of maturity level of a professional community. – 2013.
16. Dokuka S., Yavorskiy R., Krasnov F. The Structure of Organization: the Coauthorship Network Case //Analysis of Images, Social Networks and Texts. 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7-9, 2016, Revised Selected Papers. Communications in Computer and Information Science. – Springer International Publishing, 2017. – С. 93-101.
17. Guimera R. et al. Team assembly mechanisms determine collaboration network structure and team performance //Science. – 2005. – Т. 308. – №. 5722. – С. 697-702.
18. Краснов Ф.В. Модель процесса публикаций научно-практических статей по специальности 25.00 «Науки о Земле» // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 9, №5 (2017)

https://naukovedenie.ru/PDF/62TVN517.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.
19. Prell C. Social network analysis: History, theory and methodology. – Sage, 2012.
20. Brandes U. A faster algorithm for betweenness centrality //Journal of mathematical sociology. – 2001. – Т. 25. – №. 2. – С. 163-177.
21. Хасанов, Марс Магнавиевич // Википедия. [2017—2017]. Дата обновления: 16.07.2017. URL: http://ru.wikipedia.org/?oldid=86557588 (дата обращения: 16.07.2017).
22. Freeman L. C. Centrality in social networks conceptual clarification //Social networks. – 1978. – Т. 1. – №. 3. – С. 215-239.
23. Wasserman S., Faust K. Social network analysis: Methods and applications. – Cambridge university press, 1994. – Т. 8.
24. Nistér D. Preemptive RANSAC for live structure and motion estimation //Machine Vision and Applications. – 2005. – Т. 16. – №. 5. – С. 321-329.
25. Barabási A. L., Albert R. Emergence of scaling in random networks //science. – 1999. – Т. 286. – №. 5439. – С. 509-512.
26. Eom Y. H., Fortunato S. Characterizing and modeling citation dynamics //PloS one. – 2011. – Т. 6. – №. 9. – С. e24926.

Analysis of methods of construction of the graph of co-authorship: an approach based on bipartite graph

Fedor Krasnov

***Abstract* — The current practice of design and implementation of co-authorship graphs implies the use of mathematical apparatus of graph theory. Traditionally, to build co-authorship graphs using undirected graphs. The authors of this study analysed an approach of bipartite directed graph as a tool for constructing graphs of co-authorship. The study shows the benefits of using a bipartite graph and a quantitative comparison of the traditional way of constructing the graph of co-authorship and method based on a bipartite graph using the centrality metrics of the graph.**

***Keywords*— bipartite graph, co-authorship graph, centrality metrics of the graph, graph rendering, RANSAC.**