Модели поведения участников сообщества BOINC.RU

В. И. Тищенко

Институт системного анализа
Федерального исследовательского центра «Информатика и управление» Российской академии наук
117312, Москва, проспект 60-летия Октября, д.9

E-mail: vtichenko@mail.ru

В статье рассмотрено сообщество российских участников добровольных распределенных вычислений, реализуемых на программной платформе BOINC. Для проведения анализа мотивов поведения участников российского сообщества BOINC использованы данные, полученные при работе с API BOINC, приложения BOINC, и сайтом boincstats.com. Для создания базы данных российских участников был написан скрипт на PHP, хранение проводилось с помощью базы MySQL. В базе данных аккумулировались показатели по всем российским проектам, что позволило рассчитать показатели, характеризующие модели поведение российских участников.

Исходя из методологии сложных сетей, построена математическая модель сообщества BOINC.RU в виде двудольного графа с типами вершин «участник» и «проект». В качестве участников сообщества взяты аккаунты пользователей, зарегистрированных на сайте boinc.ru. Под проектами понимались зарегистрированные в системе BOINC учетные записи исследовательских проектов, в которых участники сообщества принимали участие. В гр*а*фе, представляющем сеть участников и проектов, ребро соединяет одну из вершин, относящихся к первому типу – «пользователи», а другую, ко второму – проекты, в которых пользователь, отображенный первой вершиной, принимает участие (предоставляет свои ресурсы для вычислений).

Использование методов кластеризации двудольных графов, ранее применявшихся в основном для анализа коллекций документов, позволило верифицировать основные модели поведения российских участников ДРВ при выборе и присоединении к исследовательским проектам. В отличие от описанных в литературе мотивов, выявленных методом социологического опроса участников сообщества BOINC.RU, определяющими оказываются тематические предпочтения российских участников ДРВ. И, практически, никакого влияния не оказывает на поведение участников командная или индивидуальная активность, количественно выражаемая в виде кр*е*дитов.

Полученные результаты могут иметь существенное значение при оптимизации методов управления добровольными вычислениями в сети Boinc.ru при решении задач, требующих больших вычислительных ресурсов.

Ключевые слова: добровольные распределенные вычисления, BOINC, виртуальные сообщества, сложные сети, биполярный граф, кластеризация

Введение

Сложность задач, стоящих перед современной наукой, сопровождается сложностью вычислений, необходимых для решения этих задач. При этом представление о проведение подобных вычислений обычно ассоциируется с суперкомпьютерами или вычислительными кластерами. Однако далеко не все масштабные задачи требуют такого дорогостоящего и специализированного оборудования.

С распространением и бурным развитием интернета получила практическое воплощение идея основателя проекта SETI@home («Поиска внеземного Разума) David`a Gedye – использование для масштабных распределённых вычислений ресурсов персональных компьютеров пользователей Интернета, интеграция которых обеспечена на специальной программной платформе. В качестве такой платформы используется программа BOINC[1], разработанная в 2002 году в университете Беркли [Anderson, 2003].

В основе методологии такого подхода лежит представление о сетевой организации коммуникаций и зарождении парадигмы сетевого общества в целом. Принципиальным техническим итогом решения задачи распараллеливания масштабных расчетов и разбиения исходной задачи на множество мелких, стала возможность организации («построения») вычислительных систем для добровольных распределенных вычислений (ДРВ) на принципах грид-систем.

Участники проектов ДРВ в соответствии с принципами национальной (страновой) общности, тематическими или иными предпочтениями объединяются в команды. В число 105 582 команд, объединяющих более 4 млн. участников проектов ДРВ, входит 794 российских команды. Из 54 активных проектов, использующих платформу BOINC, четыре проекта действуют на территории России: Einstein@Home, OPTIMA@home и SAT@home, Gerasim@home. Все эти проекты имеют статус «альфа»[2] [Посыпкин, 2015; Тищенко, Прочко, 2015].

Членам команд и командам в целом организаторами исследовательских проектов начисляются условные очки, так называемые, «кре́диты», количество которых зависит от предоставленных мощностей, времени участия в проектах, иных характеристик активности участников. Статистика начисления этих показателей, регулярно публикуемая на сайте www.boincstats.com, создает атмосферу состязаний, как между участниками, так и между командами.

Этот прием организаторы проектов ДРВ рассматривают в качестве одного из стимулов поддержания активности волонтеров и, соответственно, мотива сохранения их подключения (в прямом и переносном смысле слова) к проекту. В то же время, как отмечается в литературе, в целом совокупность мотивов вовлечения в проекты ДРВ пользователей интернета и «подключения» компьютеров к сети распределенных ресурсов, а также предотвращения выхода волонтеров из проектов, все еще требует своего разрешения [Darch, Carusi, 2010].

Исследование вопросов причин сотрудничества и коммуникации пользователей интернета при решении научных задач восходит к работам, анализирующим факторы формирования и развития цифровой «гражданской науки» [Nov, Arazy, …, 2011]. В этих исследованиях показано, что «категории гражданской науки», иными словами, факторы, определяющие распределение участников добровольных вычислений, зависят от характера и целей проектов. И все множество волонтеров располагается в интервале, начиная с «технических» задач, в которых их участие сводится лишь к предоставлению ресурсов компьютеров для проведения вычислений (как например, проект SETI@home или проект Folding@home компьютерного моделирования свёртывания молекул белка) и заканчивая более сложными задачами, для решения которых востребованы, как сбор, так и анализ распределенных данных (проекты Stardust@home по классификации межзвездных пылевых частиц или Galaxy Zoo построение визуальных обра-

[1] См. https://en.wikipedia.org/wiki/Berkeley_Open_Infrastructure_for_Network_Computing

[2] Статус проекта «альфа» означает, что проект функционирует и находится на начальном этапе разработки.

зов/изображений галактик). Всего, по мнению исследователей, можно описать три типа проектов и, соответственно, три распределения участников.

Выявленные различия в распределении волонтеров подчеркивают необходимость детального изучения мотивов участия в подобных проектах. И, прежде всего, в проектах ДРВ, вхождение в которые не требует от участников ничего, кроме скачивания платформы BOINC и «подключения» компьютеров к сети распределенных ресурсов. Очевидно, что реализация подобных проектов напрямую зависит от мотивов участия волонтеров в проекте, и, соответственно, от количества «подключенных» персональных компьютеров и времени предоставления их для использования в проекте. Однако у исследователей и организаторов проектов ДРВ нет однозначного ответа на вопрос – как значимый научный проект, требующий масштабных вычислений, может сформировать среду, которая будет стимулировать вклад ресурсов многими добровольцами?

По мнению исследователей BOINC-сообщества [Андреев, 2014; Якимец, Курочкин, 2015; Holohan, Garg, 2005; Nov, Arazy, …, 2014] основными мотивами участия в проектах участников VC являются:

- ощущение причастности к важным научным исследованиям;
- командный дух, переживание социального взаимодействия, идентификация с сообществом, потребность в общении с людьми, близкими по увлечениям;
- спортивный дух, атмосфера состязательности, востребованность осознания социальной статусности в виде оценок социальной активности (кре́дитов).

В основе описания этих и близких им или сопряженных по своим характеристикам мотивов лежат результаты социологических исследований участников ДРВ. И естественно они не могут не характеризоваться высокой степенью произвольности или субъективности (пусть и непроизвольных) в оценке волонтерами причин и мотивов их участия в проектах ДРВ. В этой связи важным оказывается разработка методологии верификации совокупности мотивов участников ДРВ на основе формализованных методов анализа их поведения.

Постановка задачи. Сообщество BOINC.RU как комплексная сеть

Для математической оценки взаимодействия волонтеров рассмотрим виртуальное сообщество российских участников ДРВ на платформе BOINC в форме сети. В качестве узлов сети выберем 2 типа объектов: участники сообщества (аккаунты пользователей, зарегистрированных на сайте boinc.ru) и исследовательские проекты, в которых волонтеры принимают участие (зарегистрированные в системе BOINC учетные записи проектов). В гра́фе, представляющем данную сеть, ребро соединяет одну из вершин, относящихся к первому типу объектов – пользователи, а другую, ко второму типу – проект, в котором пользователь, отображенный первой вершиной, принимает участие (предоставляет свои ресурсы для вычислений). В результате мы получим модель сети сообщества BOINC в виде двудольного графа с вершинами «участник» и «проект». Вес каждого ребра будет равен количеству «кре́дитов», заработанных участником на исследовании, с которым он связан данным ребром.

Для получения показателей характеризующих участника ДРВ в системе BOINC, количества полученных им очков, а также статистики по каждому проекту использованы сайты, на которых показатели графически визуализируются при посредстве одного из приложения BOINC – API [Программное обеспечение…, 2014].

Для проведения статистического анализа поведения российских участников ДРВ мы использовали также данные, полученные при работе с сайтом www.boinc.ru. Скрипт для получения данных и создания соответствующей базы данных с этого сайта был написан на PHP, для хранения данных использовались базы данных MySQL. В результате были получены следующие характеристики: уникальные идентификаторы участников, имена участников, уникальные идентификаторы проектов, названия проектов, количество кре́дитов у участников за послед-

нюю неделю, месяц, год и все время, принадлежность участников к проектам, уникальные идентификаторы команд, названия команд, принадлежность участников к командам.

Таким образом, была организована база данных (рис. 1), содержащая данные по пользователям, которые указали в качестве своей «принадлежности» Россию. В базе мы аккумулировали показатели по всем проектам, включая архивные, в которых российские участники ДРВ принимали участие. Это позволило рассчитать показатели, характеризующие закономерности участия российских участников в проектах и командах BOINC.

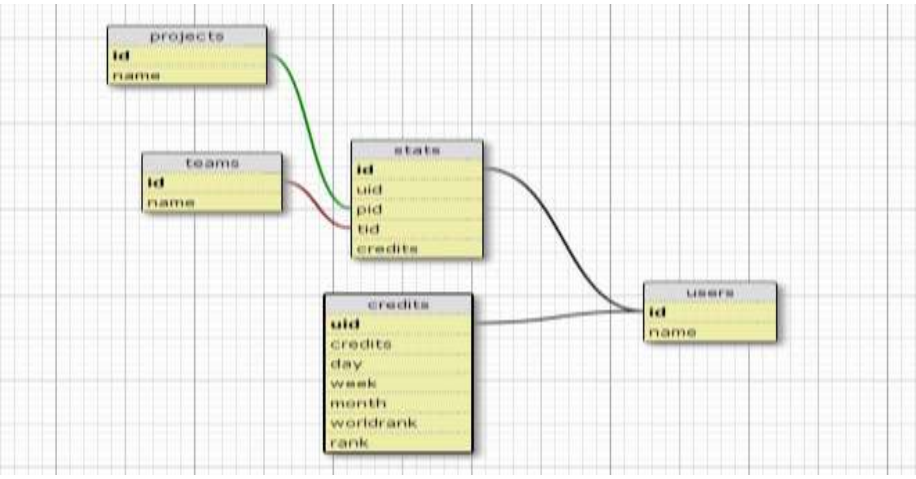

Рис. 1. Структура базы данных «Сообщество BOINC.RU»

База данных «Сообщество BOINC.RU» включает характеристики 134 проектов и 44985 российских участников. Содержащиеся в ней данные использованы при построении сети участников и проектов. Сеть была визуализирована с помощью программы Gephi. В построенном двудольном графе все показатели в сумме дали 45119 вершин, 82827 связей между ними и 740 команд. Средняя степень вершины в графе составляет примерно 1,83. Среднее количество участников в проекте – 618. Диаметр графа равен 6, средняя длина пути: 2.14. Изображения двудольного графа приведено на рисунке 2.

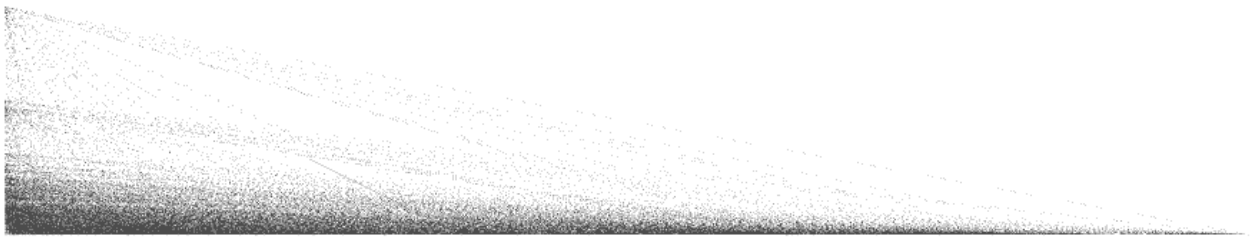

Рис. 2. На горизонтальной оси расположены проекты, на вертикальной – участники; чем дальше вершина находится от пересечения осей, тем выше ее степень (количество связей)

Анализ сети сообщества BOINC.RU методами кластеризации

Как ранее отмечалось, в качестве основных причин присоединения к проектам ДРВ исследователи на основе анализа результатов социологических опросов выделяют три группы мотивов – сопричастность к научным исследованиям, командный дух и потребность в переживании состязательности. Для верификации этих мотивов, определяющих поведение участников сообщества BOINC.RU, мы провели кластеризацию графа сообщества. По нашему мнению, резуль-

таты выделения кластеров могут показать какой из мотивов определит предпочтение волонтеров при выборе проектов – сфера научных интересов (*тематика проектов*), принадлежность к команде и участие команды в проекте (*командный дух*), оценка активности (*состязательность*), измеряемая кре́дитами.

Для оценки значимости предпочтений участников сообщества BOINC.RU были использованы четыре метода кластеризации двудольных графов:

- метод спектрального рекурсивного разбиения, Spectral Recursive Embedding (SRE) [Hongyuan, He, …, 2001];
- метод k-средних (k-means) [Dhillon, 2001];
- метод разделяющего распределения по главному направлению, Principal direction divisive partition (PDDP) [Boley, 1998];
- метод «информационного бутылочного горлышка» [Slonim, Tishby, 2000].

Алгоритмы этих методы разработаны и применялись ранее при кластеризации коллекций документов. Сравнение результатов работы методов показало высокую степень применимости каждого из них к сети участников сообщества BOINC.RU.

Из четырех примененных методов кластеризации двудольных графов по *тематической зависимости* лишь метод k-means, не показал положительных результатов. Методами SRE и PDDP алгоритмически выделены два кластера участников. Метод «информационного бутылочного горлышка» фактически полностью решил задачу поиска тематических блоков проектов и участников, распределив все множество участников по четырем тематическим кластерам. Таким образом, в результате формализованного анализа показано, что поведение участников при выборе проектов существенно зависит от их тематических предпочтений.

Предположение о влиянии *командного духа* на формирование предпочтений участников сообщества при выборе проектов, к которым они присоединяются, было подтверждено частично. При доступных вычислительных ресурсах довести дивизионные алгоритмы до итераций, на которых число кластеров сравнимо с числом реальных команд, оказалось невыполнимой задачей. Для метода с использованием алгоритма k-means предельным значением оказалось 14 средних, а для метода PDDP при 4 итерации, выделено 16 кластеров. Метод SRE оказался более пригоден и позволил проделать 5 итераций, разбив сеть на 32 кластера. Это позволило для наименьших из получившихся кластеров достичь значений количества участников в кластере, сравнимых с размерами наибольших из команд. Так из второй по численности российской команды «Russia Team», состоящей из 2837 человек, 89% (2532 участника) вошло в один из кластеров, состоящий из 5 проектов и 3873 участников. Все участники третьей по численности команды «TSC! Russia» распределились по 3 кластерам. Поскольку, такое сильное, почти полное, «вхождение» маловероятно как случайность, следует заключить, что принадлежность к командам действительно влияет на «проектную» структуру сообщества и предпочтения волонтеров.

И наконец, представление, что активность участников, оцененная в виде кре́дитов, будет сильным сигналом, влияющим на поведение при выборе проектов, не получило статистического подтверждения. В результате кластеризации графа всеми четырьмя методами, ни в одном из кластеров не обнаружилось существенного отклонения среднего количества кре́дитов для всех участников. Корреляция количества кре́дитов и количества проектов по всем участникам составила 0,23. Другими словами, их связь оказалась недостаточно сильна для заключения о каком-либо влиянии *состязательности* на предпочтения участников сообщества BOINC.RU при выборе проектов.

Заключение

Построение математической модели сообщества BOINC.RU в качестве биполярного графа позволили провести кластеризацию методами, ранее применявшимися в основном для анализа документов. В результате были верифицированы существующие представления о мотивах поведения участников ДРВ при выборе и присоединении к исследовательским проектам. В отли-

чие от описанных в литературе моделей поведения, значимыми оказываются сопричастность к научным исследованиям и социальное взаимодействие, командный дух. И, практически, никакого влияния не оказывает на поведение участников командная или индивидуальная активность, атмосфера состязательности, количественно оцениваемая в виде кре́дитов.

Полученные результаты могут иметь существенное применение при решении практических задач по управлению проектами ДРВ и оптимизации работы участников сообщества BOINC.RU.

Список литературы

Андреев А. Методы повышения популярности и привлечения участников в проектах добровольных распределенных вычислений на платформе BOINC // Национальный Суперкомпьютерный Форум (НСКФ-2014). – Переславль-Залесский, 25–27 ноября 2014 года. [Электронный ресурс]. URL: http://2014.nscf.ru/TesisAll/5_Gridi_iz_rabochix_stanciy_i_kombinirovannie_gridi/05_211_AndreevAL.pdf (дата обращения 29.10.2016).

Andreev A. Metody povysheniya populyarnosti i privlecheniya uchastnikov v proektakh dobrovol'nykh raspredelennykh vychisleniy na platforme BOINC // Natsional'nyy Superkomp'yuternyy Forum (NSKF-2014). – Pereslavl'-Zalesskiy, 25–27 November 2014. [Electronic resource]. URL: http://2014.nscf.ru/TesisAll/5_Gridi_iz_rabochix_stanciy_i_kombinirovannie_gridi/05_211_AndreevAL.pdf (accessed 29.10.2016), (In Russian).

Посыпкин М. А. Развитие технологий добровольных вычислений в России // Сборник тезисов докладов Национального суперкомпьютерного форума 23-27 ноября 2015 г. [Электронный ресурс]. URL: http://2015.nscf.ru/nauchno-prakticheskaya-konferenciya/tezisy-dokladov/ (дата обращения 27.10.2016).

Posypkin M. A. Razvitie tekhnologiy dobrovol'nykh vychisleniy v Rossii // Sbornik tezisov dokladov Natsional'nogo superkomp'yuternogo foruma 23-27 November 2015. [Electronic resource]. URL: http://2015.nscf.ru/nauchno-prakticheskaya-konferenciya/tezisy-dokladov/ (accessed 27.10.2016), (In Russian).

Программное обеспечение с открытым исходным кодом для организации добровольных распределенных вычислений и распределенных вычислений в сети. [Электронный ресурс]. URL: http://boinc.berkeley.edu/trac/wiki/GraphicsApi

The BOINC graphics API provides cross-platform support for developing graphics apps. [Electronic resource]. URL: http://boinc.berkeley.edu/trac/wiki/GraphicsApi

Тищенко В. И., Прочко А. Л. Российские участники добровольных распределенных вычислений на платформе Boinc. Статистика участия // Компьютерные исследования и моделирование. – 2015. – Т. 7, № 3. – С. 727-734.

V. I. Tishchenko, A. L. Prochko. Russian participants in BOINC-based volunteer computing projects. The activity statistics. // Computer Research and Modeling, – 2015, –Vol. 7, № 3. –PP. 727-734, (In Russian).

Якимец В. Н., Курочкин И. И. Добровольные распределенные вычисления в России: социологический анализ // Интернет и современное общество: Сборник научных статей XVIII Объединенной конференции (IMS-2015) – 2015 – Санкт-Петербург, 23-25 июня 2015 г. [Электронный ресурс]. URL: http://openbooks.ifmo.ru/ru/file/2263/2263.pdf

V. Yakimets, I. Kurochkin. Volunteer distributed computing in Russia: sociological analysis. // Internet and Modern Society – IMS: proceedings of the XVIII Joint Conference (IMS-2015) – 2015 - St. Petersburg, 23-25 June 2015. [Electronic resource]. URL: http://openbooks.ifmo.ru/ru/file/2263/2263.pdf. (In Russian).

Anderson D.P. Public Computing: Reconnecting People to Science // Presented at the Conference on Shared Knowledge and the Web, Residencia de Estudiantes, Madrid, Spain – Nov. 17-19 2003. [Electronic resource]. URL: https://boinc.berkeley.edu/madrid.html (accessed 27.10.2016).

Boley D. Principal Direction Divisive Partitioning // Data Mining and Knowledge Discovery. 1998. Vol. 2, Issue 4. P. 325 344. [Electronic resource . URL: https://www.researchgate.net/publication/227098544_Principal_Direction_Divisive_Partitioning (accessed 27.10.2016).

Darch P., Carusi A. Retaining Participants in Volunteer Computing Projects // Phil. Trans. R. Soc. A. 2010 Vol. 368. P. 4177-4192. [Electronic resource . URL: http://rsta.royalsocietypublishing.org/content/368/1926/4177 (accessed 27.10.2016).

Dhillon I.S. Co-clustering documents and words using Bipartite Spectral Graph Partitioning // Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining // Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. – 2001. – P. 269-274. [Electronic resource . URL: https://pdfs.semanticscholar.org/559d/5ca5501666ab2ad17c3fb1055d64306e9e3e.pdf (accessed 27.10.2016).

Holohan A., Garg A. Collaboration Online: The Example of Distributed Computing. // Journal of Computer-Mediated Communication. 2005. Vol. 10, Issue 4. Article 16. URL: https://www.altmetric.com/details.php?domain=onlinelibrary.wiley.com&doi=10.1111/j.1083-6101.2005.tb00279.x (accessed 27.10.2016).

Nov, O., Arazy, O., and Anderson, D. (2011). Technology-Mediated Citizen Science Participation: A Motivational Model. Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM 2011). – Barcelona, Spain, – July 2011. [Electronic resource . URL: http://faculty.poly.edu/~onov/Nov_Arazy_Anderson_Citizen_Science_ICWSM_2011.pdf (accessed 27.10.2016).

Nov O., Arazy O., Anderson D. Scientists@Home: What Drives the Quantity and Quality of Online Citizen Science Participation: PLOS One, April 1 2014. [Electronic resource URL: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0090375 (accessed 27.10.2016).

Slonim N., Tishby N. Document clustering using word clusters via the Information Bottleneck Method. In Research and development in information retrieval // Proceedings of the 23rd annual international ACM SIGIR conference. 2000. – P. 208-215. [Electronic resource . URL: http://www.cc.gatech.edu/~isbell/reading/papers/ib-word-cluster.pdf (accessed 27.10.2016).

Hongyuan Z., He. X, Ding C., Simon H., Gu M. Bipartite Graph Partitioning and Data Clustering. In: Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, November 5-10, 2001, Atlanta, Georgia, USA. P. 25-32. [Electronic resource . URL: https://arxiv.org/pdf/cs/0108018v1.pdf (accessed 27.10.2016).

Models of patterns of behavior of the participants of BOINC.RU community

V.I. Tishchenko

Institute for Systems Analysis

Federal Research Center "Computer Science and Control" of Russian Academy of Sciences,

9, prospekt 60-letya Oktyabrya, Moscow, 117312, Russia

E-mail: vtichenko@mail.ru

The article analyses the model of behavior of the Russian participants of volunteer computing (VC) using platform BOINC. The data has been received with API BOINC and site www.boincstats.com. The script for the database was written in PHP, for data storing was used MySQL. The database indicators were accumulated across all Russian projects, which allowed the calculation of the indicators characterizing the behavior of the Russian participants in all projects and teams BOINC - absolute and relative number of Russian participants, their activity, the number of introduced points system, the number of participants in each of the Russian project participants, interest in the concept of the VC.

Based on the methodology of complex networks a mathematical model of BOINC.RU community in the form of a bipartite graph with vertices types of «participants» and «project» is constructed. The participants of the community are described using accounts registered in boinc.ru site. The project – all the projects in which community members participated and registered in the system of accounts BOINC research projects. In the graph representing the network of participants and projects, edge connects one of the vertices belonging to the first type – «users» and the second – a project in which the volunteer is displayed (provides for computing resources) the first vertex participates.

By using different methods of clustering bipartite graph, previously used mainly for the analysis of collections documents, we could verify the basic models of patterns of behavior of the Russian participants of VC in selecting and joining the research projects. In contrast to the literature data based on sociological surveys our study based on clustering technique shows that the thematic preferences are deceive of the Russian participants of VC motives. And practically there is no any effect on the behavior of a team or individual activity, quantitatively expressed as crédits.

The results can be significant for optimizing VC management for solving problems that require large computational resources.

Keywords: volunteer computing, BOINC, virtual communities, complex networks, clustering, bipartite graph