ДАННЫЕ СЕРВИСА FLICKR И СТРУКТУРА СООБЩЕСТВ СТРАН

А. Б. Белый[1], Л. В. Рудикова[2], С. Л. Соболевский[3], А. Н. Курбацкий[1]

[1]*Белорусский государственный университет*
Минск, Беларусь
[2]*Учреждение образования «Гродненский государственный университет имени Янки Купалы»*
Гродно, Беларусь
[3]*Нью-Йоркский университет*
Бруклин, Соединенные Штаты Америки
e-mail: alexander.belyi@gmail.com, kurb@unibel.by, rudikowa@gmail.com, sobolevsky@nyu.edu

Рассматривается глобальная сеть перемещений людей между странами, построенная с использованием данных о цифровых гео-локализованных фотографиях и видео, размещенных в открытом доступе на интернет-сервисе Flickr. Рассматриваемый набор данных открывает новые возможности для понимания мобильности, в частности, краткосрочных поездок из одной страны в другую. В работе демонстрируется применение метода поиска сообществ к сети перемещений между странами, что позволяет выявить интересные пространственные закономерности.

Ключевые слова: сеть; мобильность; Flickr; определение сообществ.

FLICKR DATA AND COMMUNITY STRUCTURE OF THE WORLD

A. B. Belyi[1], L. V. Rudikova[2], S. L. Sobolevsky[3], A. N. Kurbatsky[1]

[1]*Belarusian State University*
Minsk, Belarus
[2]*Education Institution «Grodno State Yanka Kupala University»*
Grodno, Belarus
[3]*New York University*
Brooklyn, USA

Recent availability of geo-localized data capturing individual human activity together with the statistical data on international migration opened up unprecedented opportunities for a study on global mobility. In this paper we consider it from the perspective of a complex network, built using a dataset of digital photos and videos posted on the Flickr website. This dataset provides insights on the global mobility highlighting short-term visits of people from one country to another. We use this mobility network to infer the structure of the global society through a community detection approach and demonstrate that consideration of mobility network between countries can reveal interesting global spatial patterns.

Keywords: network; mobility; Flickr; definition of community.

ВВЕДЕНИЕ

Все чаще, путешествуя из одной страны в другую, люди оставляют за собой цифровой след в различного рода сервисах. Соответствующие данные открывают огромные возможности для исследований: с их помощью мы можем восстановить перемещения людей, проанализировать их и, возможно, обнаружить интересные и важные закономерности. Рассматривая мобильность в глобальном масштабе, крайне важно учитывать различные аспекты человеческих перемещений, которые состоят из разных видов мобильности и включают как релокацию на постоянное место жительства, так и краткосрочные посещения. Анализ данных из различных источников позволяет отдельно рассматривать международные миграции и краткосрочные путешествия.

Предыдущие исследования миграционных данных и данных сервиса Twitter показали, что определение сообществ в сетях мобильности и взаимодействий людей обычно приводит к географически связным сообществам (даже при том, что никакие пространственные характеристики методом определения сообществ не учитываются), выявляя важные географические закономерности. В предлагаемой работе применен аналогичный подход к сети, основанной на данных сервиса Flickr.

Сеть перемещений между странами, полученная из данных сервиса Flickr, в основном представляет краткосрочные перемещения, так как в большинстве случаев данные Flickr отражают туристическую активность. В статье исследуется топология этой сети, ее структура сообществ, т. е. разбиение стран на кластеры. Выделение структуры сообществ и понимание ее связи со спецификой стран является крайне важным с точки зрения международного туризма. Действительно, нахождение сообществ в сети мобильности означает определение кластеров стран, которые имеют тесно переплетенные культурные и исторические связи между ними, будучи относительно менее тесно связанными со странами вне кластера.

В статье использована максимизация модулярности для того, чтобы определить разбиение. Однако, чтобы учесть отсутствие петель в сети, проведена корректировка выражения для функции модулярности. В описании результатов дается обсуждение разбиения на различных уровнях гранулярности, найденных путем кластеризации с различными значениями параметра разрешения.

НАБОР ДАННЫХ

Используемый набор данных Flickr состоит из 130 миллионов фотографий и видеофайлов. Он был собран из двух наборов, размещенных в открытом доступе [1–2]. Набор содержит данные за десять лет: с 2005 по 2014 г. Используя указанные данные, построена ориентированная взвешенная сеть краткосрочных перемещений, в которой вершины соответствуют странам, а веса ребер равны количеству пользователей из одной страны, посетивших другую страну. Для этого первоначально определена постоянная страна пребывания пользователей (когда это было возможно). Затем, если пользователь имел фотографии или видео, сделанные в других странах, это трактовалось как посещение им данной страны. Для определения страны постоянного пребывания использован один из наиболее консервативных методов, применяемых в подобных исследованиях [3]: такой страной считается та, в которой пользователь сделал наибольшее количество фотографий (не менее 10) и провел наибольшее количество времени (не менее полугода). Используя этот критерий, стало возможным определить

страну постоянного пребывания для более чем 500 тыс. пользователей, которые сделали более 80 % всех фотографий и видео, т. е. более 90 млн. При исследовании также исключены из рассмотрения страны, которые соответствуют вершинам с входящей или исходящей силой менее 10. В результате исследования получена сеть из 201 страны.

ОПРЕДЕЛЕНИЕ СООБЩЕСТВ

Для определения сообществ использован один из наиболее популярных и хорошо установившихся подходов к разбиению сетей, основанный на максимизации функции модулярности. Но поскольку в рассматриваемой сети отсутствуют петли, проведена корректировка классической формулы. В частности, изменен способ, которым нуль-модель, используемая в модулярности, оценивает ожидаемый вес каждого ребра. В своей классической форме модулярность использует $\frac{s_i t_j}{\sum_k s_k}$ как ожидаемый вес ребра из начальной вершины i в конечную вершину j, где s_i и t_j – входящая и исходящая сила вершин i и j соответственно, и m – суммарный вес всех ребер сети, т. е. $s_i = \sum_j w_{ij}$, $t_j = \sum_i w_{ij}$ и $m = \sum_k s_k = \sum_k t_k = \sum_{ij} w_{ij}$, в то время как w_{ij} – наблюдаемый вес ребра из i в j. Одно из возможных объяснений такого ожидаемого значения ($\frac{s_i t_j}{m} = \frac{s_i t_j}{\sum_k t_k}$): в случайной сети, которая сохраняла бы силы вершин, распределение исходящей силы i и s_i между всеми возможными конечными вершинами должно быть пропорционально входящим силам t_j этих конечных вершин. Другое – поскольку $\frac{s_i t_j}{m} = \frac{s_i t_j}{\sum_k s_k}$, распределение входящих сил t_j между всеми возможными начальными вершинами должно быть пропорционально исходящим силам s_i этих начальных вершин. Однако если петли не участвуют в этом распределении, то ожидаемый вес скорее должен быть равен $\frac{s_i t_j}{\sum_{k \neq i} s_k}$ или $\frac{s_i t_j}{\sum_{k \neq j} t_k}$ в зависимости от того, рассматривать его как распределение исходящей силы s_i между всеми конечными вершинами, за исключением самой i, или как распределение входящей силы t_j между всеми начальными вершинами, за исключением самой j. В качестве окончательного варианта использовано среднее этих двух значений, что привело к выражению $\frac{1}{2}\left(\frac{s_i t_j}{m - t_i} + \frac{s_i t_j}{m - s_j}\right)$.

Поскольку известно, что модулярность имеет определенные недостатки, такие как, например, предел разрешающей способности, не позволяющий определять достаточно мелкие сообщества, также использован подход, предложенный ранее в [4], который вводит так называемый параметр разрешающей способности. Это привело к дальнейшим изменениям в формуле для модулярности. Таким образом, окончательная формула для скорректированной модулярности, используемая в рассматриваемом случае для сети без петель, выглядит следующим образом:

$$Q = \frac{1}{2m} \sum_{i \neq j} \left(2w_{ij} - a\frac{s_i t_j}{m - t_i} - a\frac{s_i t_j}{m - s_j}\right) \delta(C_i, C_j),$$

где a – параметр разрешающей способности, i, j – вершины, C_i, C_j – сообщества, которым они принадлежат, $\delta(x, y) = 1$, если $x = y$, иначе 0, Q – модулярность.

Наконец, для нахождения наилучшего разбиения оптимизирована и рассматриваемая версия модулярности, используя точный и эффективный алгоритм Combo [5], в целях максимизации различных типов целевых функций.

СТРУКТУРА СООБЩЕСТВ СЕТИ ПЕРЕМЕЩЕНИЙ МЕЖДУ СТРАНАМИ

Максимизация модулярности при значении разрешающего параметра 1,0 приводит к выявлению *пяти* сообществ, в то время как для значения 2,0 число найденных сообществ возрастает до *семнадцати*, делая рассмотрение больших значений затруднительным, поскольку становится сложно различить визуально и анализировать разные сообщества на карте. На рис. 1 и 2 приведены разбиения для значений разрешающего параметра 1,0 и 2,0 соответственно; страны, попавшие в одно сообщество, окрашены одним цветом.

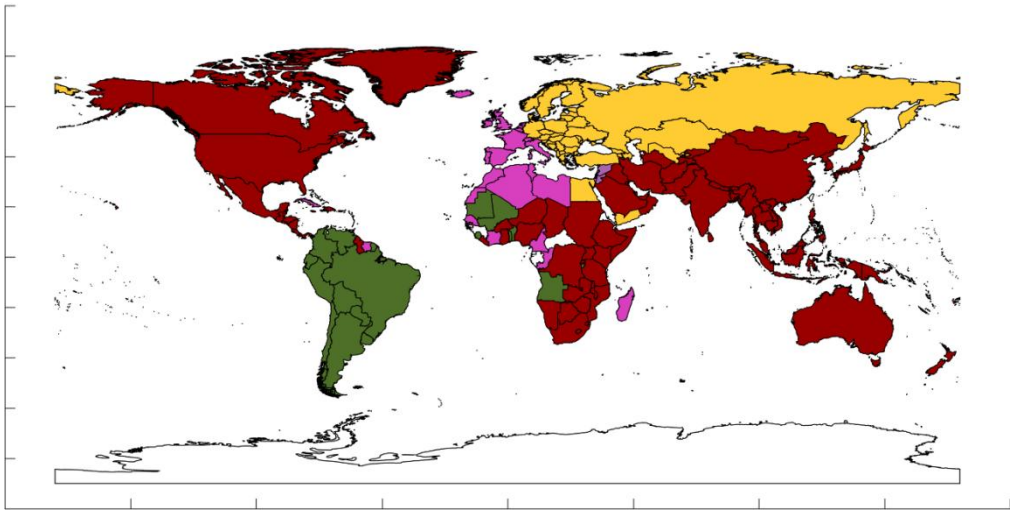

Рис. 1. Разбиение стран на сообщества при a = 1,0

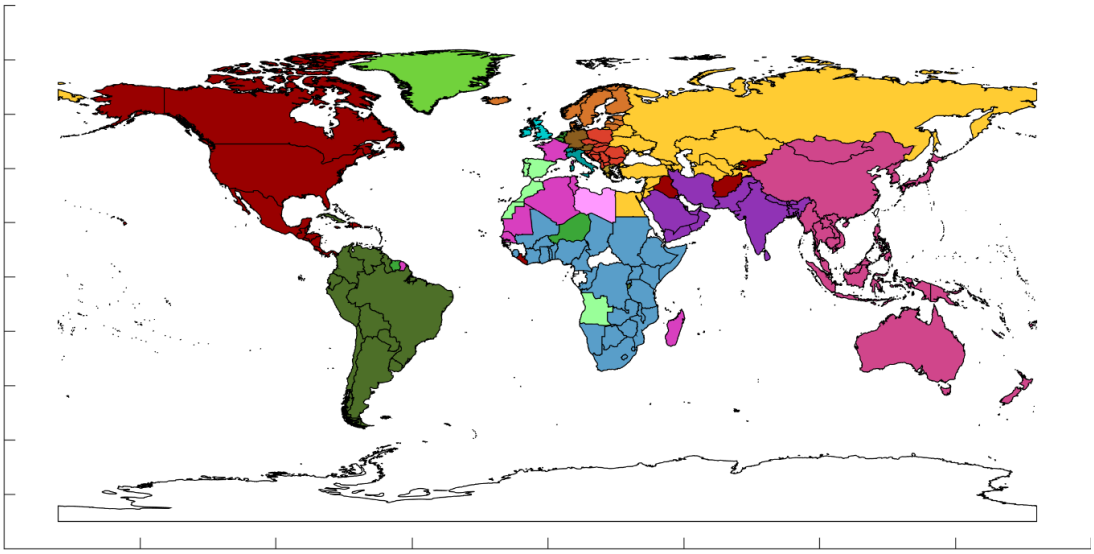

Рис. 2. Разбиение стран на сообщества при a = 2,0

Из рисунков видно, что основные географические регионы, такие как Северная и Южная Америка, Восточная Азия, страны СНГ, объединены в отдельные сообщества. Для разных значений параметра можно наблюдать интересные особенности. Напри-

мер, Египет и Турция попадают в одно сообщество со странами СНГ, что может быть объяснено их популярностью среди туристов из СНГ. Если рассматривать структуру сообществ большей гранулярности (рис. 2), то можно заметить сильную связь Ирака и Афганистана со странами Северной Америки, как и европейских стран с их бывшими африканскими колониями. В то же время только Ирландия попадает в одно сообщество с Великобританией, некогда могущественным доминионом с колониями по всему миру. Полученные кластеры указывают на то, что хотя люди чаще путешествуют в близлежащие страны, общий язык и история играют важную роль в выборе страны назначения очередного путешествия.

Также интересным результатом является то, что структура сообществ в сети мобильности, построенной по данным сервиса Flickr, обладает свойством, присущим большинству других изученных сетей мобильности: сообщества географически связаны и отражают устоявшиеся регионы. Это согласуется с предыдущими работами, посвященными исследованию сетей мобильности, построенными при помощи использования данных телефонных звонков, сервиса Twitter и миграций. Причины для объединения стран в одно сообщество могут включать близкое географическое положение, культурные аспекты и сильные экономические связи.

ЗАКЛЮЧЕНИЕ

В предлагаемой работе исследована сеть краткосрочных перемещений людей, созданная с использованием набора данных о медиа-объектах, размещенных в интернет-сервисе Flickr. Для изучения структуры мирового сообщества к рассматриваемой сети применен алгоритм определения сообществ. В результате исследования выявлены специфические пространственные закономерности, а также подтверждено отмеченное в предыдущих работах наблюдение, что разбиения сетей мобильности представляют собой установившиеся географически связанные сообщества.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. SFGEO.ORG [Electronic resource]. URL: http://sfgeo.org/data/tourist-local (date of access: 02.09.2016).

2. YFCC100M: The New Data in Multimedia Research / B. Thomee [et al.] // Communications of the ACM. 2016. Vol. 59, № 2. P. 64–73.

3. Choosing the Right Home Location Definition Method for the Given Dataset / I. Bojic [et al.] // Social Informatics. Springer International Publishing, 2015. P. 194–208.

4. Arenas A., Fernandez V., Gomez S. Analysis of the structure of complex networks at different resolution levels // New J. of Phys. 2008. Vol. 10. P. 053039.

5. General optimization technique for high-quality community detection in complex networks / S. Sobolevsky [et al.] // Phys. Rev. E. 2014. Vol. 90, № 1. P. 012811.