ДОКЛАДЫ АКАДЕМИИ НАУК, 2018, том 481, №1, с. 10-14

МАТЕМАТИКА

УДК 512.643

КЛАСТЕРНЫЙ КОЭФФИЦИЕНТ В МОДЕЛИ ПРОСТРАНСТВЕННОГО ПРЕДПОЧТИТЕЛЬНОГО ПРИСОЕДИНЕНИЯ

© 2018 Л. Исхаков[1,*], М. Миронов[1], Л. Прохоренкова[1,2], Б. Каминski (B. Kaminski)[3], П. Пралат (P. Pralat)[4]

Представлено академиком РАН В.В. Козловым 14.02.2018 г.

Поступило 22.02.2018 г.

Рассматривается кластерная структура графа в модели пространственного предпочтительного соединения, для которой уже было показано сходство с реальными сетями во многих аспектах. Изучается поведение локального кластерного коэффициента, а именно, рассматривается асимптотическое поведение его среднего значения по всем вершинам графа некоторой степени при стремлении размера графа к бесконечности. Данная величина не была рассмотрена ранее и отражает типичную зависимость кластерной структуры в окрестности некоторой вершины от её степени в графе. Кроме того, показано наличие с высокой вероятностью вершины, для которой значение коэффициента отличается от его среднего значения.

В последнее время в научной литературе большое внимание уделяется сложным сетям. Исследования показывают, что сети, полученные из реальных данных, таких как социальные сети и др., имеют ряд типичных свойств, например малый диаметр, степенное распределение степеней вершин, кластерную структуру [10]. Для анализа и предсказания свойств растущих с каждым днём сетей было предложено достаточно много математических моделей случайных графов.

Так, например, диаметр случайного дистанционного графа частично изучен в работе [3]. Часто также моделирование сложных сетей используют для анализа социальных сетей. В работе [2] представлен результат о времени смешивания марковских цепей при построении искусственных социальных графов.

Наиболее изученным свойством сложных сетей является распределение их степеней вершин. Для большинства реальных сетей это распределение хорошо приближается степенным распределением с параметром γ, который обычно находится в интервале (2, 3) [11]. Другое важное свойство реальных сетей – их кластерная структура. Одним из способов её измерения является *кластерный коэффициент*. В некотором смысле его значение есть вероятность для двух соседей одной вершины быть смежными между собой. В литературе предложено два классических определения: средний локальный кластерный коэффициент и глобальный кластерный коэффициент (формальные определения даны в разделе 2). Было показано, что во многих реальных сетях оба этих коэффициента стремятся к положительной константе с увеличением размера сети [10].

Предметом данного исследования является кластерная структура модели *пространственного предпочтительного присоединения* (*Spatial Preferential Attachment*, SPA), которая была предложена в работе [1]. Эта модель естественным образом сочетает в себе предпочтительное присоединение и геометрическую составляющую. Особенно нас будет интересовать средний локальный кластерный коэффициент $C(d)$ как функция от степени d вершины. Модель SPA хорошо изучена, было показано, что она схожа с реальными сетями во многих аспектах. Например, в [1] доказан степенной закон распределения степеней вершин. Кластерный коэффициент $C(d)$ для этой модели не был изучен, хотя некоторые кластерные свойства проанализированы в [4]. В работах [4] и [5] доказано, что средний локальный кластерный коэффициент сходится по вероятности к положительной константе, если и только если распределение степеней вершин имеет конечную дисперсию.

1. Модель SPA

1.1. Определения

Модель SPA комбинирует предпочтительное присоединение и структуру метрического

1) Лаборатория продвинутой комбинаторики и сетевых приложений, Московский физико-технический институт (государственный университет), Долгопрудный
2) Управление машинного интеллекта и исследований, Яндекс, Москва, Россия
3) Warsaw School of Economics, Warsaw, Poland
4) Ryerson University, Toronto, Canada
*E-mail: lenar-iskhakov@yandex.ru

пространства, в котором лежат вершины. Достигается это за счёт наличия у каждой вершины “сферы влияния”. Параметрами модели являются вероятность создания ребра $p \in [0,1]$ и две константы *A1, A2*, где

$$0 < A_1 < \tfrac{1}{p},\ A_2 > 0,$$

которые отвечают за объём сфер влияния вершин. Вершины в данной модели – это точки m-мерного единичного гиперкуба $S = [0,1]^m$ с торической метрикой на нём, индуцированной нормой L_k, т.е.

$$d(x,y) = \min\{\| x - y + u \|_k : u \in \{-1,0,1\}^m\} \quad \forall x, y \in S.$$

В модели SPA генерируется последовательность случайных графов $\{G_t\}$, где $G_t = (V_t, E_t)$, $V_t \subseteq S$. Пусть $\deg^-(v,t)$ – входящая степень вершины v в графе G_t, и $\deg^+(v,t)$ – её исходящая степень. Тогда «сфера влияния» $S(v,t)$ вершины v в момент времени $t \geq 1$ – это шар с центром в v и объёма:

$$|S(v,t)| = \min\left\{\frac{A_1 \deg^-(v,t) + A_2}{t}, 1\right\}.$$

Последовательность графов строится постепенно, начиная с пустого графа G_0 в момент времени t = 0. Далее для каждого $t > 1$ граф G_t получается из графа G_{t-1} следующим образом. Сначала случайно и равномерно из гиперкуба S выбирается новая вершина v_t, которая добавляется к множеству вершин V_{t-1}. Затем независимо для каждой такой вершины $u \in V_{t-1}$, что $v_t \in S(u, t-1)$, направленное ребро (v_t, u) проводится с вероятностью p. Таким образом, вероятность того, что ребро (v_t, u) будет добавлено в момент времени t, равна $p|S(u, t-1)|$.

На рис. 1 представлен граф в SPA-модели.

1.2. Свойства модели

В этом разделе мы приведём некоторые изученные ранее свойства и приложения SPA-модели. Теорема 1.1 из [1] утверждает, что SPA-модель генерирует граф со степенным распределением входящих степеней вершин с параметром $1 + 1/(pA_1)$. С другой стороны, средняя исходящая степень асимптотически равна $pA_2/(1 - pA_1)$. Теорема 1.3. В [6] показано, что SPA-модель хорошо приближает графовую структуру социальных графов, полученных из сети Facebook. Обычно предполагается, что распределение вершин в гиперкубе S равномерно [7], однако в [8] рассматриваются и неравномерные распределения, что более реалистично для некоторых приложений модели.

2. Кластерный коэффициент

Кластерный коэффициент измеряет, какова вероятность того, что два соседа одной вершины соединены ребром. В литературе предложено несколько определений кластерного коэффициента.

Глобальный кластерный коэффициент $C_{\text{glob}}(G)$ – отношение утроенного количества

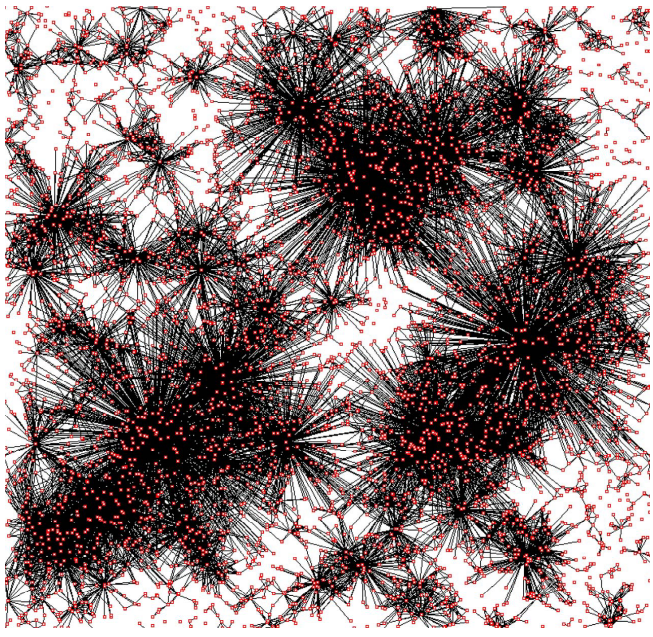

Рис. 1 Граф в SPA модели при *t = 500, p = 1 b A1 = A2 = 1*

треугольников к количеству пар смежных рёбер в графе G. Другими словами, если мы выберем случайную пару смежных рёбер, то $C_{\text{glob}}(G)$ будет вероятностью того, что три данные вершины образуют треугольник. Глобальный кластерный коэффициент в модели SPA рассматривался в [4, 5]. Мы же, в свою очередь, рассмотрим *локальный кластерный коэффициент*. Определим его для неориентированного графа $G=(V,E)$. Пусть $N(v)$ − множество соседей вершины v, $|N(v)|=\deg(v)$. Для любого $B\subseteq V$ пусть $E(B)$ − множество рёбер в подграфе, индуцированном на множество B:

$$E(B)=\{(u,w)\in E: u,w\in B\}.$$

Тогда *кластерный коэффициент* вершины v определяется следующим образом:

$$\frac{c(v)=|E(N(v))|}{\binom{\deg(v)}{2}}.$$

Видно, что $0\le c(v)\le 1$. Усреднением по всем вершинам степени d получаем $C(d)$ − средний локальный коэффициент по вершинам степени d:

$$C(d)=\frac{\sum_{v:\deg(v)=d} c(v)}{|\{v:\deg(v)=d\}|}.$$

Далее, мы будем использовать обозначения $c(v,t)$ и $C(d,t)$, соответствующие графам в момент времени t.

Локальный кластерный коэффициент $C(d)$ в различных сетях и их моделях был широко изучен как с теоретической, так и с эмпирической точки зрения. Например, было показано, что в реальных сетях $C(d)$ обычно убывает как $d^{-\psi}$ для некоторого $\psi>0$. В частности, [14] показано, что $C(d)$ может быть хорошо приближен с помощью d^{-1} для некоторых больших сетей, а в [15] получено степенное убывание $C(d)$ с параметром 0,75. Локальный кластерный коэффициент был изучен в ряде вероятностных моделей сложных сетей. Например, в [9] показано, что для некоторых таких моделей $C(d)\sim d^{-1}$. Мы получили схожее поведение и для модели SPA.

Напомним, что SPA-модель генерирует ориентированный граф, поэтому формально определим локальный кластерный коэффициент для ориентированного случая. Пусть $N^-(v,t)\subseteq V_t$ − множество вершин, из которых ведёт ребро в вершину v в момент времени t. Тогда интересующая нас величина определяется как

$$\frac{c^-(v,t)=|E(N^-(v,t))|}{\binom{\deg^-(v,t)}{2}}.$$

Как и в неориентированном случае, определим

$$C^-(d,t)=\frac{\sum_{v:\deg^-(v,t)=d} c^-(v,t)}{|\{v:\deg^-(v,t)=d\}|}.$$

3. Результаты

Все результаты, которые будут представлены в этом разделе, носят асимптотический характер. Это значит, что некоторое свойство последовательности графов $\{G_n\}$ рассматривается при n, стремящемся к бесконечности. Мы говорим, что событие выполнено *асимптотически почти наверное* (а.п.н.), если оно выполнено с вероятностью, стремящейся к единице при $n\to\infty$. Также для множества S мы говорим, что *почти все* элементы из S обладают некоторым свойством P, если число элементов S, не удовлетворяющих свойству P, есть $o(|S|)$. Наконец, мы используем обозначение $f\ll g$ *для* $f=o(g)$ и $f\gg g$ *для* $g=o(f)$.

Как было сказано ранее, мы хотим показать, что SPA-модель схожа с реальными большими сетями в терминах локального кластерного коэффициента. А именно, мы покажем, что данный коэффициент в SPA-модели для вершин степени k может быть хорошо приближен функцией $\frac{1}{k}$ при $n\to\infty$.

Мы рассмотрим кластерный коэффициент для ориентированного графа, однако стоит иметь в виду, что аналогичные утверждения верны и для соответствующего неориентированного графа, который получен из исходного заменой всех направленных рёбер на ненаправленные.

Для SPA-модели мы можем не только оценить кластерный коэффициент $C^-(d,n)$, но и сформулировать некоторое вероятностное утверждение для отдельно взятой вершины v. Однако для этого её входящая степень должна быть довольно большой.

Изложение начнём с теоремы, которая показывает негативный для нашей цели результат. А именно, если некоторая вершина v итоговой входящей степени k оказалась в зоне плотного скопления более ранних вершин, то её кластерный коэффициент может оказаться довольно большим (значительно больше, чем k^{-1}). Из этого следует, что утверждение о том, что для всех без исключения

вершин в графе локальный кластерный коэффициент такой, как мы ожидаем, неверно.

Теорема 1. Пусть

$$C = 5\ln\left(\frac{1}{p}\right)$$

и

$$\xi = \xi(n) = \frac{1}{(\omega(\ln\ln n)^2(\ln\ln\ln n))} = o(1)$$

для некоторого $\omega = \omega(n)$, *стремящегося к бесконечности при* $n \to \infty$.

Предположим, что $k = k(n)$ *и* $2 \le k \le n^{\xi}$.

Тогда а.п.н. существует такая вершина v, *что* $\deg^-(v,n) \sim k$ *и выполнено*

(i) $c^-(v,n) = 1$, *при условии* $2 \le k \le \sqrt{\frac{\ln n}{C}}$

(ii) $c^-(v,n) = \Omega(1) \gg \frac{1}{k}$, *при условии*

$$\sqrt{\frac{\ln n}{C}} \le k \le \frac{\ln n}{\ln\ln n}$$

(iii) $$c^-(v,n) \gg \frac{(\ln\ln n)^2(\ln\ln\ln nn)}{k} \gg \frac{1}{k}$$

при условии

$$\frac{\ln n}{\ln\ln n} \le k \le n^{\xi}$$

.

Однако несмотря на это, почти все вершины достаточно большой степени имеют кластерный коэффициент порядка $1/k$, *что сформулировано в следующей теореме.*

Теорема 2. Пусть $\varepsilon, \delta \in (0, \frac{1}{2})$ *некоторые константы, а* $k = k(n) \le n^{pA_1 - \varepsilon}$. *Пусть* $-$ *подмножество вершин графа* G_n, *в котором каждая вершина имеет входящую степень между* $(1-\delta)k$ *и* $(1+\delta)k$. *Тогда а.п.н. выполнено следующее:*

(i) почти все вершины из X_k *имеют*

$c^-(v,n) = \Theta(\frac{1}{k})$, *при условии* $k \gg \ln^{C_1} n$

где

$$C_1 = 4 + \frac{(4pA_1 + 2)}{(pA_1(1 - pA_1))} .$$

(ii) средний кластерный коэффициент $c^-(v, n)$ *вершин из* X_1 *есть* $\Theta(\frac{1}{k})$, *т.е.*

$$\frac{1}{|X_k|}\sum_{v \in X_k} c^-(v,n) = \Theta(\frac{1}{k}),$$

при условии $k \gg \ln^{C_2} n$, *где*

$$C_2 = 4 + \frac{(4pA_1 + 2)}{(pA_1(1 - pA_1))} .$$

Приведённые результаты показывают схожесть модели SPA и реальных графов в терминах локального кластерного коэффициента, а также дают подробное представление о значении этого коэффициента в различных случаях.

СПИСОК ЛИТЕРАТУРЫ

1. *Aiello W., Bonato A., Cooper C., Janssen J., Pralat P.* A Spatial Web Graph Model with Local Influence Regions. Internet Math. 2009. V. 5. P. 175-196.
2. *Avrachenkov K., Iskhakov L., Mironov M.* On Mixing in Pairwise Markov Random Fields with Application to Social Networks. In Proc. XIII Intern. Workshop WAW. 2016. Montreal, QC, Canada, December 14-15. 2016. p.127–139. Montreal. Springer, 2016.
3. *Iskhakov L., Mironov M.* Diameters of Random Distance Graphs. J. of Math. Sci.. 2017. №227(4). P. 407-418.
4. *Jacob E., Morters, P. A.* Spatial Preferential Attachment Model with Local Clustering. In International Workshop on Algorithms and Models for the Web-Graph. B: Springer. 2013. P. 14-25.
5. *Jacob E., Morters P., et al.* Spatial Preferential Attachment Networks: Power Laws and Clustering Coefficients. //The Ann. App. Probab. 2015. №25(2). P. 632-662.
6. *Janssen J., Hurshman M., Kalyaniwalla N.* Model Selection for Social Networks Using Graphlets//Internet Mathematics. 2013. V. 8. №4. P. 338-363.
7. *Janssen J., Pralat P., Wilson, R.* Geometric Graph Properties of the Spatial Preferred Attachment Model. //Adv. in App. Math. 2013. V.50. P. 243-267.
8. *Janssen J., Pralat P., Wilson R.* Non-uniform Distribution of Nodes in the Spatial Preferential Attachment Model. //Internet Mathematics. 2016. V. 12. №1-2. P. 121-144.
9. *Krot A. and Prokhorenkova L. O.* Local Clustering Coeficient in generalized Preferential Attachment Models//In: Intern.Workshop on Algorithms and Models for the Web-Graph. B: Springer. 2015. P. 15-28.
10. *Newman M. E.* The structure and Function of Complex Networks. //SIAM review, 2003, v. 45(2): p.167–256.
11. *Newman M. E.* Power Laws, Pareto Distributions and Zipf's Law//Contemp. physics. 2005. V. 46. №5. P. 323-351.
12. *Prokhorenkova L.O.* General Results on Preferential Attachment and Clustering

Coeficient//Optim. Letters. 2017. V. 11. №2. P. 279-298.
13. Raigorodskii A. Small Subgraphs in Preferential Attachment Networks.//Optim. Letters. 2017. V. 11. №2. P. 249-257.
14. Ravasz E., Barabasi A.-L. Hierarchical Organization in Complex Networks//Phys. Rev. E. 2003. V. 67. №2: 026112.
15. Vazquez A., Pastor-Satorras R., Vespignani A., Large-scale Topological and Dynamical Properties of the Internet//Phys. Rev. E. 2002. V. 65. №6: 066130.