Анализ показателей центральности с заданным разбиением на сообщества с использованием линейной пороговой модели

Гарбуз Владислав

Декабрь 2023

Содержание

1 Введение

Выявление наиболее влиятельных узлов в сложных сетях имеет большое практическое применение и актуальность в реальном мире, так как направленные на них действия могут позволить значительно укрепить или ослабить распространение информации (будь то слухи или маркетинговая кампания) или же эпидемий (посредсвом помощи в определении стратегии вакцинации). Также они применяются и в других сферах, взаимодействие элементов в которых можно представить в виде рёбер, а сами элементы вершинами. Примерами таких систем могут служить социальные сети, компьютерные сети, электросети, нейронные сети и так далее. Однако в действительности стоит учитывать, что в подобных сетях зачастую можно выделить сообщества — подмножества вершин, внутри которых связи между вершинами плотные, а за пределами, между различными такими сообществами, менее плотные. Оказывается, что меры центральности, использующиеся для выявления наиболее важынх узлов, что учитывают разбиение графа на сообщества, выгодно отличаются от классических мер, которые эту информацию не используют.

В данной работе проводится анализ специфики каждой из тем, с последующим сравнением популярных показателей центральности, учитывающих сообщества. Ядро обзора, главным образом, составляет статья "Analyzing Community-Aware Centrality Measures Using the Linear Threshold Model" за авторством Stephany Rajeh, Ali Yassin, Ali Jaber, and Hocine Cherifi, посвящённая построению линейной пороговой модели, с целью изучения эффективности различных мер центральностей, учитывающих сообщества.

2 Анализ содержания статей

В начале аргументируем утверждение о практической пользе и актуальности применения сложных сетей для явлений реального мира. В комплексной статье Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications сотрудники университета Сан-Паулу изучают подобную интеграцию: они проанализировали более 20 областей, среди которых встречаются нейронауки, лингвистика, экология, транспорт, коммуникации и многие другие, и пришли к выводу, что, благодаря своей универсальности и гибкости, сложные сети уже активно применяются в огромном количестве сфер, начиная от неврологии и заканчивая землетрясениями, для решения прикладных задач. Более того, наиболее востребованными сложные сети оказались в решении различных задач, связанных с белками (авторам удалось обнаружить порядка 50 возможных приложений), в то время как изначальной мотивацией для изучения сложных сетей являлись возможные приложения в области Интернета, которых обнаружилось меньше — лишь 42.

Отдельно можно отметить статьи Weighted Adaptive Neighborhood Hypergraph Partitioning for Image Segmentation и A combinatorial edge detection algorithm on noisy images за авторством Soufiane Rital, Hocine Cherifi, Serge Miguet и Soufiane Rital, A. Bretto, Hocine Cherifi, Driss Aboutajdine соответсвенно. Первая статья является продолжением второй и обе посвящены частному приложению сетей для обнаружения границ объектов на зашумленных изображениях, путём сопоставления пикселей изображения вершинам гиперграфа и использованием дальнейших алгоритмов, учитывающих взвешенные связи вершин (пикселей).

Следующим шагом следует рассмотреть процесс выделения вершин графа в сообщества. Этому посвящена статья Community structure in social and biological networks, написанная Michelle Girvan и Mark E. J. Newman. В своей работе они предлагают свой метод выделения сообществ внутри сетей, который был протестирован ими на графах, структура сообществ которых зарнее известна. В результате их метод демонстрирует выявление этой структуры с высокой чувствительностью и надежностью. В основе работы их метода, традиционным образом, находится иерархическая кластеризация, которая базируется на построении дендрограммы. Для этого для каждой пары вершин вычисляется некоторый вес, который может высчитываться различными способами, самый простой из которых — степень посредничества. Это мера измеряет количество кратчайших путей между всеми парами узлов, которые проходят через данное ребро. Далее строится сама дендрограмма: все вершины пред-

ставляются листьями между которыми начинают строить рёбра в порядке их весов, начиная с пары с наибольшим весом и продвигаясь к самой слабой. Таким образом, "Срез" этого дерева на любом из уровней даёт сообщества, которые существовали перед добавлением ребра соответствующего веса. Однако предложенный метод отличается от остальных тем, что рёбра соединяются, а, наоборот, удаляет рёбра с высоким весом, отчего граф начинает распадаться на те самые сообщества.

Выделение на сообщества позволяет перейти к обзору наиболее популярных показателей центральности, учитывающих сообщества. Первая из них — Comm Centrality, алгоритм и формула которой подробно описаны в статье Centrality Measures for Networks with Community Structure, за авторством Naveen Gupta, Anurag Singh и Hocine Cherif. Обозначим формулу:

$$CC(i) = (1+\mu_C)\cdot\left(\frac{k^i_{\text{intra}}}{\max_j(k^j_{\text{intra}}\forall j\in C)}\cdot R\right)+(1-\mu_C)\cdot\left(\frac{k^i_{\text{inter}}}{\max_j(k^j_{\text{inter}}\forall j\in C)}\cdot R\right)^2$$

$$\text{где } \mu_C = \frac{\text{Количество внешних связей в сообществе C}}{\text{Общее количество связей в сообществе C}}$$

Опишем переменные: μ_C — доля внешних связей в сообществе C. k^i_{intra} — количество внутренних связей узла i в сообществе C. k^i_{inter} — количество внешних связей узла i из сообщества C. $\max_j(k^j_{\text{intra/inter}}\forall j \in C)$ — максимальное количество внутренних/внешних связей узла в сообществе C. R — коэффициент масштабирования, нормализующий значения внутренних и внешних связей.

Итак, мера центральности Comm Centrality учитывает важность узла внутри своего сообщества как и существенность для связи с другими сообществами. Для вычисления этой меры используется комбинация количества связей узла внутри своего сообщества (внутренних связей) и связей с узлами в других сообществах (внешних связей). Узел, обладающий множеством связей внутри своего сообщества, можно считать "лидером" или "хабом" внутри этого сообщества. Помимо этого, если узел также имеет много связей с узлами в других сообществах, то его можно рассматривать как "мост" или "связующее звено" между разными сообществами. Также стоит отметить, что мостам придаётся больший вес. Таким образом, данная мера центральности пытается найти узлы, которые одновременно являются важными внутри своего сообщества, но в то же время играют важную роль связующих элементов между разными сообществами в сети.

Обсудим следующую меру центральности — Community-Based Centrality. Она была предложена и подробно описана в статье A Community-Based Approach to Identifying Influential Spreaders сотрудниками/учащимеся китайских университетов Zhiying Zhao, Xiaofan Wang, Xiaofan Wang и Zhiliang Zhu. Вкратце, предложенный показатель центральности объединяет количество внутриобщинных и межобщинных связей, взвешенных по размеру их сообществ. Авторы отмечают, что введённый ими показатель обладает малой вычислительной сложностью, а также отмечают, что, при прочих равных, узел с более высоким показателем Community-Based Centrality оказывает большее влияние на распространение по сравнению с узлом с большой степенью посреднечества, обсуждённой ранее в обзоре. Кроме того, введённая мера также обладает стабильностью в оценке важности узлов при вариации параметров, что подчёркивает её эффективность в анализе распространения в сложных сетях. Приведём формулу:

$$\alpha_{CBC}(i) = \frac{1}{N} \sum_{c=1}^{N_c} \frac{k_{i,c}}{n_c}$$

где N_c - общее количество сообществ в сети. $k_{i,c}$ — количество связей узла i с узлами в сообществе c. n_c - количество узлов в сообществе c. N — общее количество узлов в графе.

Третьей мерой центральности является Community-Based Mediator, введённую в статье Identifying Influential Nodes Based on Community Structure to Speed up the Dissemination of Information in Complex Network за авторством Muluneh Mekonnen Tulu, Ronghui Hou, Talha Younas. Данная метрика измеряет важность узла посредством учитывания энтропии случайного блуждания от узла к каждому сообществу в сети. Она описывает, насколько узел является ключевым для соединения двух или более сообществ в сети. Напишем формулу:

$$CBM_i = \left(-\sum \rho_{intra_i} \log(\rho_{intra_i}) - \sum \rho_{inter_i} \log(\rho_{inter_i}) \right) \times \frac{k_i}{\sum_{i=1}^{N} k_i}$$

где $\rho_{intra_i} = \frac{k_{intra_i}}{k_i}$ вес внутренних связей для узла i и $\rho_{inter_i} = \frac{k_{inter_i}}{k_i}$ вес внешних связей для узла i.

Важным замечанием является слова авторов о том, что CBM оказывается эффективнее в сравнении с традиционными методами, такими как степень посредничества, но, что более интересно, эффективнее своего предшетсвенника в лице CBC.

Так происходит потому что узлы, выбранные с использованием Community-Based Mediator, являются наиболее промежуточными узлами (главными "посредниками"), которые получают и передают информацию в сети лучше, чем другие узлы. Как следствие, результаты симуляций, приведённых в статье, демонстрируют, что узел с высоким значением CBM оказывает большее воздействие на распространение информации в сети, чем узел с высокой степенью CBC.

Четвёртой рассмотренной мерой центральности будет Community Hub-Bridge. Она подробно описывается в статье Immunization of networks with non-overlapping community structure, написанной Zakariya Ghalmane, Mohammed El Hassouni и Hocine Cherif. Простыми словами, мера Community Hub-Bridge основана на сочетании количества связей внутри сообщества, взвешенных по размеру сообщества, и связей между сообществами, взвешенных по количеству соседних сообществ. То есть CHB пытается одновременно выделять узлы, являющиеся ключевыми для внутренней координации внутри сообщества, а также являющиеся важными в связывании различных сообществ в единую сеть. Отсюда и происходит название, включающее в себя упоминание хабов и мостов. Однако в исследовании отмечается, что такая мера работает хорошо главным образом только на графах с чётко выраженными сообществами, в противном случае она демонстрирует плохие результаты. Предъявим формулу:

$$\alpha_{CHB(i)} = |c_q| \times k_{\text{intra}_i} + \sum_{C_l \subset C \setminus c_q} \bigvee_{j \in C_l} a_{ij} \times k_{\text{inter}}$$

где $|c_q|$ — количество узлов в сообществе c_q. k_{intra_i} — количество внутренних связей узла i. k_{inter} — количество внешних связей между узлом i и другими узлами в сообществе C_l. Наконец, $\bigvee_{j \in C_l} a_{ij}$ — обозначает логическое ИЛИ по всем j в C_l, где $a_{ij} = 1$ если узел i соединён хотя бы с одним узлом j в сообществе C_l.

Следующий показатель центральности использует понятие модулярности. Модулярность — это мера структуры сетей для измерения силы разбиения графа на сообщества. С увеличением значения модулярности, повышается плотность связи между узлами внутри сообществ, и ослабевают связи между узлами в различных сообществах. Свежая статья Measuring Node Contribution to Community Structure With Modularity Vitality, созданная Thomas Magelinski, Mihovil Bartulovic и Kathleen M. Carley, посвящена введению меры центральности, основанной на модулярности. Modularity Vitality оценивает вклад узла путём его удаления и измерения изменения модульности. Такой подход позволяет более полно оценить важность каждой

вершины: метод не только определяет, насколько важен узел для целостности своего сообщества, но и указывает, является ли узел мостом между сообществами или центром внутри своего сообщества. Это позволяет лучше понять роль узлов во всей структуре сети и их влияние на коммуникацию и распространение информации в сообществах. Другими словами, данный подход скорее можно обозначить приставкой макро-, в то время как остальным больше подходит микро-. Обозначим формулу:

$$\alpha_{MV}(i) = |M(G_i) - M(G)|$$

где M это модулярность графа, а $M(G_i)$ модулярность после удаления узла i.

Следующая статья, Functional Cartography of Complex Metabolic Networks, написанная Roger Guimera, Luís A Nunes Amaral, рассматривает показатель центральности Participation Coefficient в контексте классификации узлов (метаболитов) в биологических сетях. Данная мера центральности принимает значение от 0 до 1, где близость к 1 указывает на то, что связи узла равномерно распределены между всеми сообществами, в то время как близость к 0 означает, что все связи узла находятся в пределах только его собственного сообщества. Выходит, что Participation Coefficient позволяет оценить важность узла как посредника между различными сообществами. Это важно для понимания структуры и функциональности сети, поскольку узлы с высоким значением данной меры центральности могут иметь важное значение для поддержания связности и взаимодействия между различными частями сети. Укажем формулу с использованием уже привычных обозначений:

$$\alpha_{PC}(i) = 1 - \sum_{c=1}^{N_c} \left(\frac{k_{i,c_q}}{k_i} \right)^2$$

Последним показателем центральности, рассмотренным в настоящем обзоре станет K-Shell with Community, описанный в статье Identifying Influential Spreaders of Epidemics on Community Networks за авторством Shi-Long Luo, Kai Gong, Li Kang. Мера основана на нлокальном и глобальном влиянии узла иерархической декомпозиции k-оболочек, где оба влияния взвешиваются по определяемому пользователем параметру δ, установленному в данном исследовании равным 0,5. Понятие k-оболочки подразумевает под собой уровень или слой внутри сети, содержащий узлы с одинаковым или похожим числом связей. То есть узлы внутри одной k-оболочки имеют схожую или одинаковую степень, которая определяется как число связей вершины. Обобщая, заключаем, что понятие k-оболочки предоставляет нам структурирован-

ный способ рассмотрения сети, выделяя уровни связности узлов в зависимости от числа их связей. Параметр δ в формуле помогает распределить приоритет между внутренними и внешними связями (относительно сообществ). Формула:

$$\alpha_{ks}(i) = \delta \cdot \beta_{intra}(i) + (1 - \delta) \cdot \beta_{inter}(i)$$

где $\beta_{intra}(i)$ и $\beta_{inter}(i)$ это значение k-оболочки узла i с учетом только внутри-сообщественных и между-сообщественных связей, соответственно.

После завершения обсуждения мер центральностей (учитывающих разбиение на сообщества), мы можем перейти к описанию линейнуй пороговой модели, введённой в статье от 1978 года под названием Threshold Models of Collective Behavior, написанной Mark Granovetter. В ней рассматривается модель поведения коллектива с использованием концепции порогов. Для этого проводится исследование, как индивиды в группе принимают решения, основанные на действиях своих соседей, и как эти решения могут привести к коллективному поведению. Самое главное заключается в следующем: у каждого индивида есть свой уровень порога, который представляет собой количество его соседей, необходимое для того, чтобы он изменил своё поведение. В случае, когда число соседей, поддерживающих определенное действие, превышает порог индивида, то он также начинает поддерживать это действие. Подобное моделирование помогает понять, как индивидуальные решения могут привести к широкомасштабным изменениям в поведении группы, что отлично ложится на нашу итогувую цель — сравнить эффективность рассматриваемых мер центральности.

Обратимся к главной статье, посвящённой нашей цели — Analyzing Community-Aware Centrality Measures Using the Linear Threshold Model, написанной Stephany Rajeh, Ali Yassin, Ali Jaber, Hocine Cherifi. Зададим формальное описание пороговой модели для нашей цели, для этого процитируем часть статьи: "Рассмотрим граф $G = (V, E)$, в котором $(u, v) \in E$ представляет собой ребро между узлами u и v. Каждый узел v обладает порогом $\theta_v \in [0, 1]$, и в любой момент времени t узел v может находиться только в одном из двух альтернативных состояний:

$$v_i(t) = \begin{cases} 0, & \text{если } v_i \text{ неактивен}, \\ 1, & \text{если } v_i \text{ активен}. \end{cases}$$

Изначально доля X узлов активна. На каждом временном шаге неактивный узел v проверяет состояния своих соседей. Он активируется, если количество активных

соседей m_v удовлетворяет условию:

$$\frac{m_v}{k_v} \geq \theta_v,$$

где k_v - степень узла v. Процесс повторяется до тех пор, пока не будет активировано больше узлов. Активный узел остается активным до завершения процесса диффузии."

Дальше авторы применяют линейную пороговую модель на 13 имеющихся реальных сетях, 4 из которых представляют социальные сети (например, страницы политиков Facebook), три биологические сети (например, Human Protein), три инфраструктурные сети (например, энергосистема США) и три сети сотрудничества (GrQc, AstroPh, DBLP), с различными порогами: средний $\theta_1 = 0.4$, высоким $\theta_2 = 0.7$ и с случайным порогом, равномерно распределённом на отрезке от 0 до 1 $\theta_3 = U[0, 1]$. Из построенной визуализации исследователи делают следующие выводы: для среднего порога первые два места разделяют между собой Comm Centrality и Modularity Vitality, за ними идёт Community-based Mediator; в лучае высокого порога происходят небольшие изменения — Community-based Mediator занимает первое место, за ним идёт Comm Centrality, а далее Modularity Vitality; для случайного порога нет строгого победителя, однако первые три места явно принадлежат уже упомянутым Comm Centrality, Community-based Mediator и Modularity Vitality. Все остальные показатели центральности, учитывающие разбиение на сообщества, показывают себя существенно хуже. Community Hub-Bridge же уверенно показывает одни из самых худших показателей, что можно было предсказать, учитывая рассмотренный ранее недостаток, связанный с плохой эффективностью на графах с слабым выделением сообществ.

3 Критическое сравнение, оценка недостатков и преимуществ статей, а также сделанных в них допущений

Статьи, посвящённые мерам центральности отличались особенно содержательными визуализациями, в особенности это было присуще статьям Measuring node contribution to community structure with modularity vitality и Identifying influential nodes based on community structure to speed up the dissemination of information in complex network. Тем же преимуществом обладает и статья Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications, что содержит наиболее подробное описание процесса исследования.

Один из самых больших недостатков можно обнаружить в статье Community structure in social and biological networks. Дело в том, что описывая алгоритм выделения сообществ на графе при помощи иерархической кластеризации, авторы не указывают применяемую меру между кластерами (сообществами), которых насчитывается более пяти, что делает воспроизводимость исследования затруднительным. Также авторы центральной статьи Analyzing Community-Aware Centrality Measures Using the Linear Threshold Model зачастую прибегают к крайне скудному описанию первоначальной настройке исследования, что вновь пагубно отражается на воспроизводимости результатов.

4 Заключение

Выявление наиболее влиятельных узлов в сложных сетях действительно имеет большой потенциал в решении различных задач в самом разном спектре областей. Более того, подобные методы находят применение уже сейчас в массе приложений. Также, нами было исследовано, что сложные сети, представляющие реальные данных, зачастую обладают свойством разделения на сообщества, своего рода отдельные узлы с превосходящей внутри плотностью связей, по сравнению с наружными рёбрами. Далее было рассмотрено 7 различных мер центральностей, учитывающих разбиения сетей на сообщества, а именно их концепции и принципы работы. В заключительной части было рассмотренно исследование посвящённое поиску наиболее эффективной меры центральности, из описанных, при помощи линейной пороговой модели. Фаворитами стали Comm Centrality, Community-based Mediator и Modularity Vitality — они продемонстрировали лучшие результаты на череде тестов с использованием различных порогов и доли изначально заражённых узлов. Вершины заражались в порядке убывания значимости, определяемой избранной мерой центральности, но не превосходя изначально заданной доли. При активации в первую очередь вершин, имеющих наибольшую степень важности, по трём названным выше мерам, заражение происходило ощутимо быстрее, чем при использовании других показателей центральности, что демонстрирует их надёжную работу.

5 Источники

1. Stephany Rajeh, Ali Yassin, Ali Jaber, Hocine Cherifi: Analyzing Community-Aware Centrality Measures Using the Linear Threshold Model, January 2022.

2. Luciano da Fontoura Costa and others: Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications, September 2008.

3. Soufiane Rital, Hocine Cherifi, Serge Miguet: Weighted Adaptive Neighborhood Hypergraph Partitioning for Image Segmentation, August 2005.

4. Soufiane Rital, A. Bretto, Hocine Cherifi, Driss Aboutajdine: A combinatorial edge detection algorithm on noisy images, February 2002.

5. Michelle Girvan, Mark E. J. Newman: Community structure in social and biological networks, Juny 2002.

6. Naveen Gupta, Anurag Singh, Hocine Cherif:Centrality Measures for Networks with Community Structure, February 2016.

7. Zhiying Zhao, Xiaofan Wang, Xiaofan Wang, Zhiliang Zhu: A Community-Based Approach to Identifying Influential Spreaders, April 2015.

8. Muluneh Mekonnen Tulu, Ronghui Hou, Talha Younas: Identifying Influential Nodes Based on Community Structure to Speed up the Dissemination of Information in Complex Network, January 2018.

9. Zakariya Ghalmane, Mohammed El Hassouni, Hocine Cherif: Immunization of networks with non-overlapping community structure, June 2018.

10. Thomas Magelinski, Mihovil Bartulovic, Kathleen M. Carley: Measuring Node Contribution to Community Structure With Modularity Vitality, January 2021.

11. Roger Guimera, Luís A Nunes Amaral: Functional Cartography of Complex Metabolic Networks, January 2021.

12. Shi-Long Luo, Kai Gong, Li Kang: Identifying Influential Spreaders of Epidemics on Community Networks, Jan 2016.

13. Mark Granovetter: Threshold Models of Collective Behavior, May 1978.