Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра вычислительных методов

Курсовая работа

Информационные каскады в социальных сетях
со сложной передачей сообщений

Работу выполнил:
Д. А. Питеркин
Группа 304
Научный руководитель:
В. А. Шведовский

2021

Содержание

1. Введение

С ростом популярности общения в интернете анализ социальных сетей представляет все больший интерес для социологов, экономистов, политологов, математиков, специалистов по *computer science*, а также для практического применения в политике, бизнесе. Многие популярные онлайн-сервисы представляют собой практическую реализацию математической социальной сети, и с их помощью можно получать удобные для изучения данные.

Модель социальной сети — это граф, в котором вершины являются **агентами** (ими могут быть люди и другие субъекты, способные передавать информацию — например, новостной канал, публичная страница или политическая партия), а ребра выражают связи агентов — их способность передавать друг другу информацию. Ребра могут быть направленными и ненаправленными. В направленной сети сообщение может передаваться только по направлению ребра. Связанные ребром агенты называются **соседями**.

Большой интерес представляет тема передачи сообщений (**message passing**) в социальных сетях. Она затрагивает такие вопросы, как диффузия инноваций в обществе, информационные каскады и другие. Многие модели, несмотря на свою простоту, нашли эмпирическое доказательство и полезность в реальной жизни. В настоящей работе предлагается модель, призванная исправить недостатки существующих моделей.

Информационный каскад (**information cascade**) — явление в социальной сети, когда агенты начинают принимать решения на основе не только собственной информации, но и от наблюдаемого поведения других агентов. Понятие близко к феномену группового мышления, стадного инстинкта.

2. Примеры моделей диффузии инноваций

В большинстве моделей диффузии инноваций агенты в начальный отсчет времени делятся на уже обладающих информацией (**seeds**) и на еще не обладающих информацией (идеей, мнением, продуктом). С течением времени сообщение передается через агентов дальше по сети. Это очень напоминает модель развития эпидемии болезни (например, модель SI с зараженными и здоровыми людьми). Диффузию инноваций в социальных сетях также часто называют социальным заражением (**social contagion**). Можно вспомнить, что видеоролики, стремительно распространяющиеся в интернете, называют вирусными. Далее для простоты тоже будем называть **зараженными** агентов, обладающих рассматриваемой идеей.

Самая простая модель передачи сообщений в социальной сети — пороговая линейная модель (**linear threshold model**). В ней каждый агент обладает параметром **порога**, а каждое ребро обладает параметром **влияния**. Незараженный агент может стать зараженным, только если он соседствует с достаточным количеством зараженных агентов, и входящие ребра суммарно обладают влиянием, превосходящим порог. Итеративно вся социальная сеть может стать зараженной или остановиться на каком-то моменте:

Если $\sum w_{ij} \geq W_i$, то агент i становится зараженным. w_{ij} — входящие веса влияний соседних агентов, W_i — порог агента i.

Другая популярная модель — независимые каскады (**independent cascades**). В ней

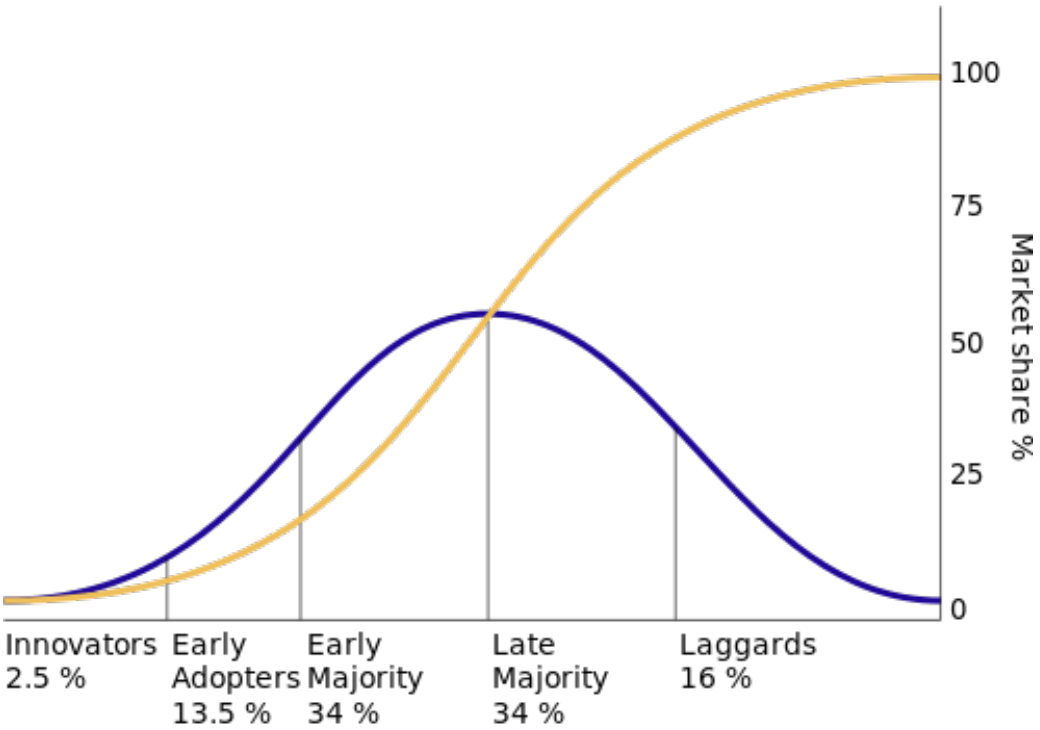
каждый зараженный агент i имеет шанс один раз с некоторой вероятностью p_{ij} ($0 \leq p_{ij} \leq 1$) заразить своего соседа j. Существует разновидность модели, когда шансов дается сколько угодно. В таком случае, если агенты имеют ненулевой шанс передать сообщение, вся сеть рано или поздно точно станет зараженной, при условии, что все ребра между агентами являются ненаправленными. Если агент имеет одинаковый шанс заразить любого соседа, то его обозначают параметром вершины p_i и называют влиянием агента.

Социальные сети также исследуют с помощью неграфовых моделей. Модель диффузии Басса (**Bass diffusion model**) — модель диффузии инноваций в виде динамической системы:

$$\frac{dF(t)}{dt} = (p + qF(t))(1 - F(t))$$

где $F(t)$ — количество принявших инновацию людей, p — коэффициент инновации, q — коэффициент имитации. Большой q позволяет соседям быстро перенимать инновацию от общности других людей. Коэффициент p выражает возможность человека получить идею вне рассматриваемой социальной сети. Например, когда начинается дождь, каждый человек открывает зонтик независимо от соседей.

При анализе передачи сообщения в сети возникает задача определить, сколько времени займет процесс передачи и как его оптимизировать, создать ситуацию, чтобы каскад не затух. В обоих случаях хороший способ — задать в качестве начальных зараженных самых влиятельных и/или самых соединенных агентов. Выгоднее распространять свой продукт, заказав рекламу у популярного источника, а не у случайного человека. Для этого у каждого агента можно определить его **центральность** — меру соединенности с остальной сетью. Существуют разные методы это сделать, самый простой — просто подсчитать количество соседей.



(рис. 1) Процесс принятия инновации со временем

Процесс принятия инновации обществом представляет собой сигмоиду, что подтверждает большинство моделей и реальные данные. Людей, принявших инновацию к определенному моменту времени, разделяют на инноваторов, ранних перенимателей инновации, раннее и позднее большинство, запаздывающих.

Перечисленные модели являются слишком упрощенными версиями реального мира. Они не предусматривают явления реактивного сопротивления инновациям, плохо объясняют, как небольшие подсообщества могут генерировать большие информационные каскады, а дорогая реклама, наоборот, часто неудачна и не генерирует информационные каскады.

С развитием интернета и сферы больших данных появляется возможность создавать более сложные модели с большим количеством настраиваемых параметров.

3. Предлагаемая модель

Попытаемся более детально изобразить распространение отдельного сообщения по социальной сети. Как известно, идею можно выразить разными способами. Например, реклама может быть шуточной или серьезной. От того, в каком именно формате будет решено распространять продукт, зависит его вирусная успешность.

Была выдвинута гипотеза, что возможно подобрать параметры для сообщения таким образом, чтобы оно смогло создать максимально возможный информационный каскад. Этого можно добиться анализом рассматриваемой социальной сети.

Рассмотрим направленный граф:

- У каждого ребра, соединяющей соседей, имеется векторный параметр w_{ij} длиной n
- У каждой вершины имеется вещественный параметр b_i
- Передаваемое сообщение обладает векторным параметром x длиной n

Параметр x описывает формат, в котором зараженный агент будет пытаться передавать его соседям. Компоненты вектора x ограничим от -1 (противоположность смыслу компоненты параметра) до 1 (полное соответствие).
Параметр w_{ij} описывает то, как передаваемое сообщение воспримется принимающей стороной.
Параметр b_i описывает общее влияние агента.

При передаче сообщения для каждой исходящие ребра от зараженного агента к незараженному вычисляем величину $p_{ij} = \sigma(w_{ij} \cdot x + b_i)$, где $\sigma(x) = \frac{1}{1+e^{-x}}$ — сигмоида.
p_{ij} — это вероятность незараженного агента на следующем шаге стать зараженной, $0 < p_{ij} < 1$.

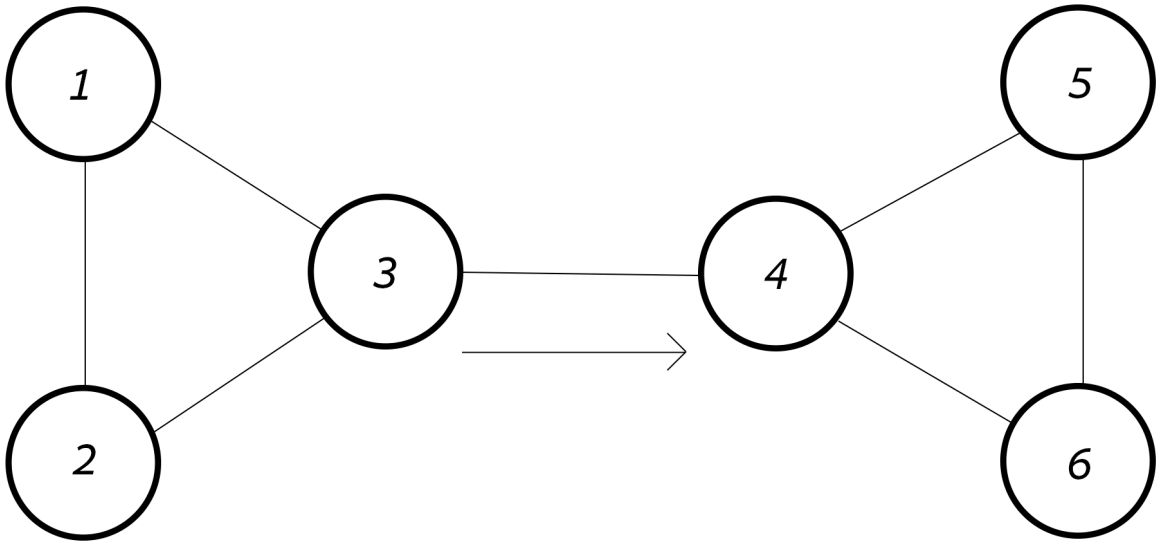
Пример:

- Компания решила рекламировать продукт в шуточной форме. Пусть его параметр единичной длины $x = (1)$, где 1 выражает большую долю юмора в сообщении.
- Реклама была куплена у популярного телевизионного канала с влиянием $b_0 = 5$

- Недавно в стране прошел траур, поэтому все шутки людьми сейчас воспринимаются плохо. Пусть исходящие из телевизионного канала ребра до зрителей имеют параметр $w_{0j} = (-10)$, что обозначает неприязнь, реактивное сопротивление к шуткам.

- Тогда все $p_{0j} = \sigma(-5) \approx 0.0067$. Если реакция других агентов на сообщения между собой схожая, то с такими шансами сообщение будет распространяться очень долго, особенно учитывая, что влияние у них меньше, чем у канала.

В некоторых более сложных случаях с более длинным вектором-параметром, одни негативные черты сообщения могут быть скомпенсированы положительными, несколько небольших положительных могут в комбинации дать большой вклад. Эти и другие возможные комбинации выражают процессы в реальной жизни, и их можно эксплуатировать для создания больших информационных каскадов.



(рис. 2) Передача сообщения между двумя сообществами по мосту

Если сообщества в сети соединены мостом (небольшим количеством смежных агентов), то взаимоотношения агентов непосредственно на мосту имеют решающую роль при информационном каскаде.

В реальной жизни общество действительно состоит из огромного количества подсообществ. Поэтому можно предположить, что описанная модель лучше описывает реальные процессы по сравнению с простейшими моделями из предыдущего раздела.

4. Вычислительные эксперименты

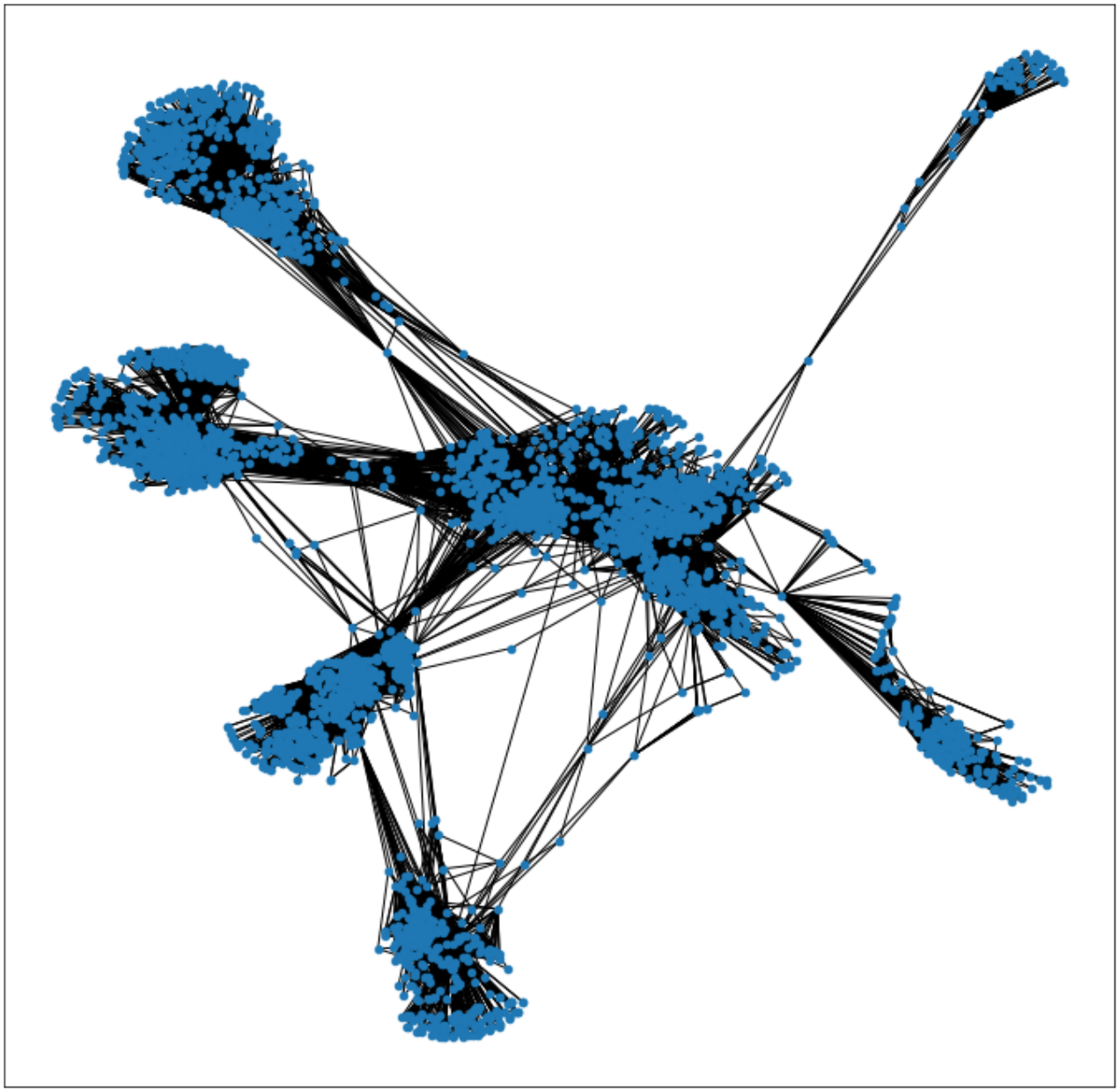
Рассматриваемую модель, как и любую агентную модель, затруднительно исследовать аналитически, поэтому было решено провести серию вычислительных экспериментов с

различными параметрами, чтобы выявить закономерности.

В исследовании был использован язык программирования Python. С кодом исследования в виде Jupyter Notebook можно ознакомиться на Github: https://github.com/MitPitt/coursework. Для реализации графов и операций на них была использована библиотека NetworkX, для визуализации — библиотека matplotlib.

В качестве примера социальной сети взят набор данных "Social circles: Facebook" (https://snap.stanford.edu/data/egonets-Facebook.html) Стэнфордского университета, являющийся реальным отрезком сети Facebook. В нем 4039 вершин и 88234 ребра.

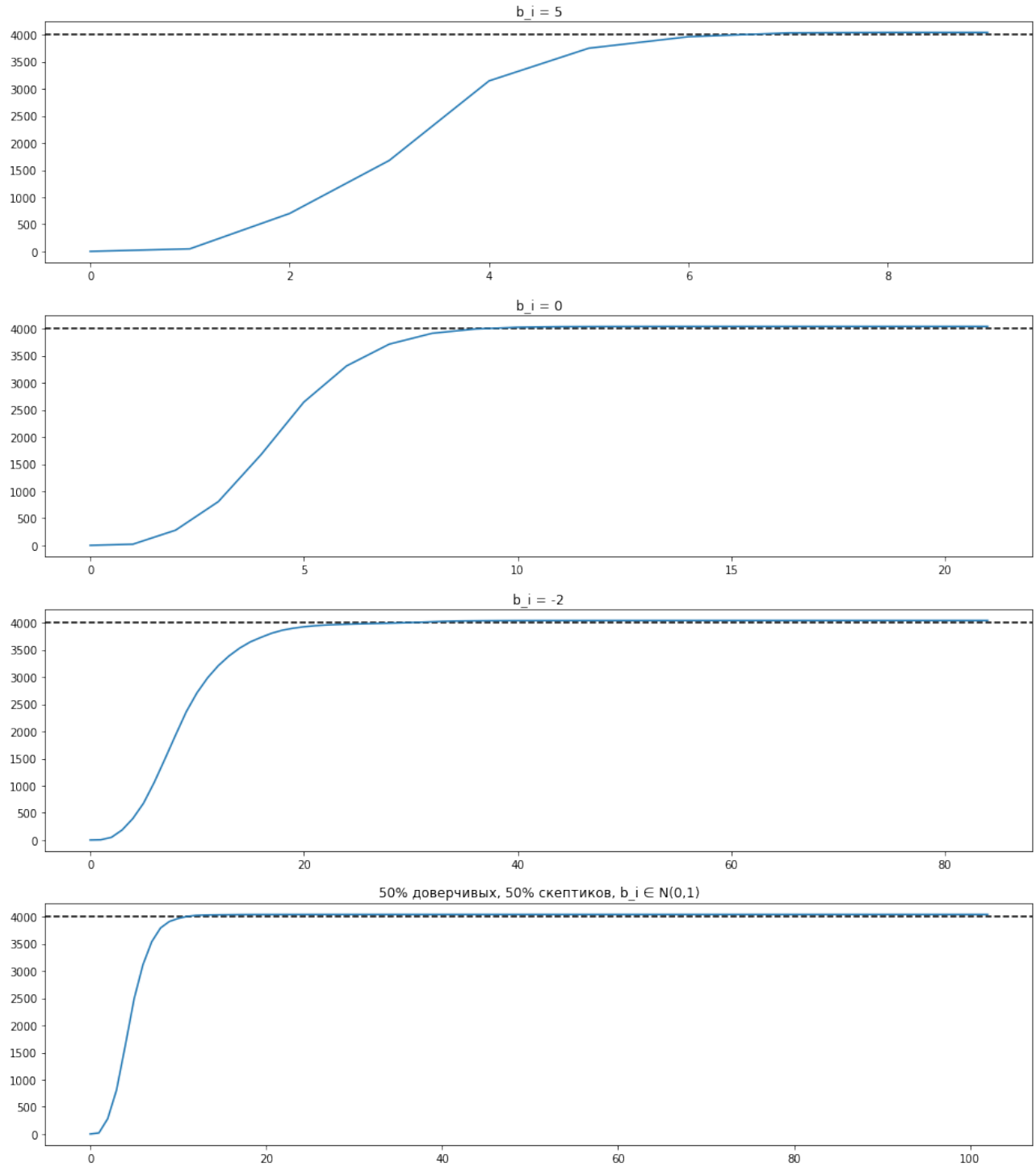
Все эксперименты с заданными параметрами вычисляются 100 раз со случайными начальными зараженными агентами. Затем результаты усредняются и выводятся на графике, отражающем процесс полного заражения сети со временем. В силу неполноты используемых данных, ограничимся ненаправленным графом, где соседи имеют возможность заразить друг друга.



(рис. 3) Визуализация социальной сети в наборе данных "Social circles: Facebook"

4.1. Независимые каскады

Если взять нулевую длину вектора x сообщения, то действующим остается только параметр влияния b_i, и получается модель независимых каскадов: $p_{ij} = \sigma(b_i)$. Рассмотрим примеры с различными b_i:



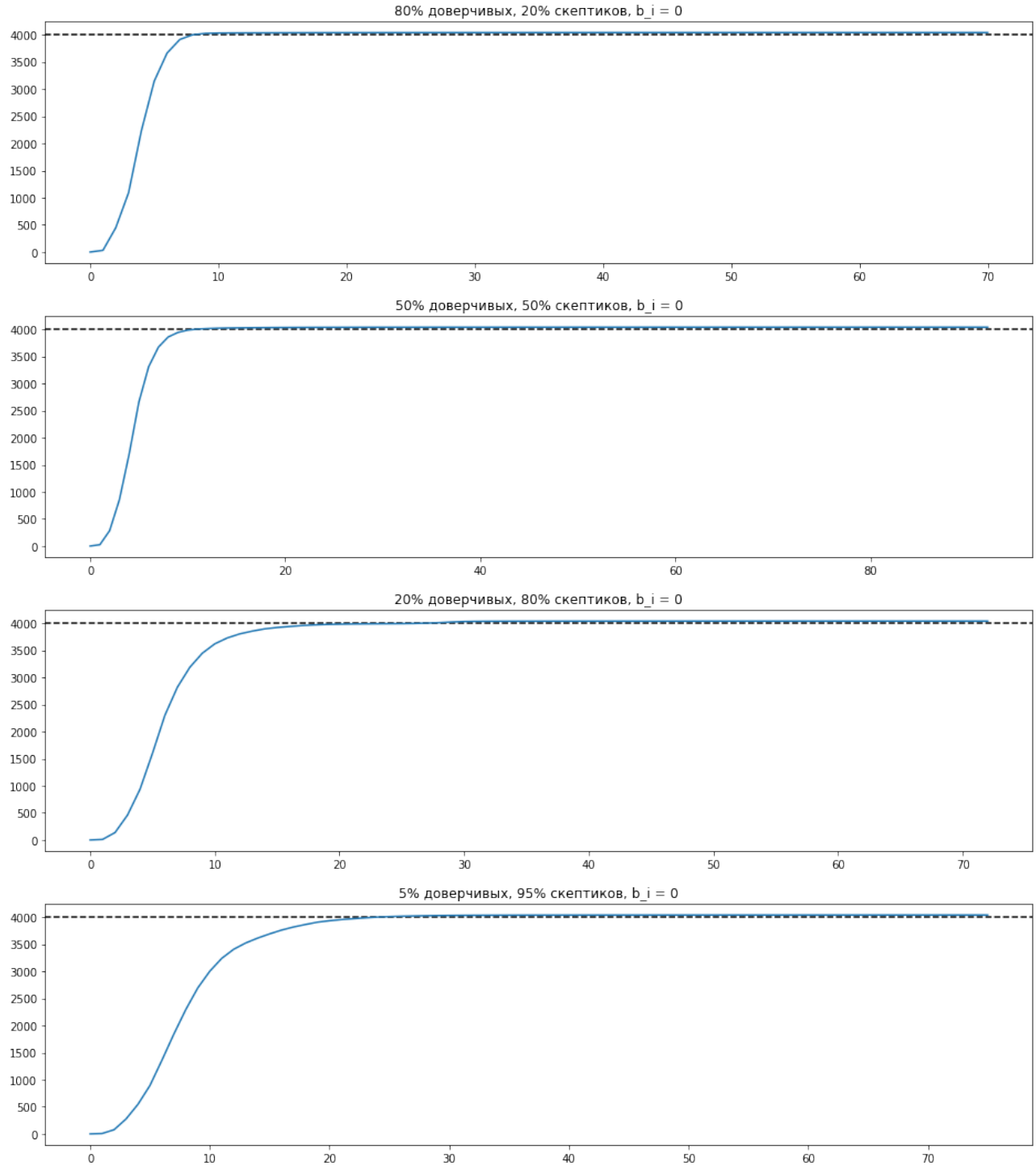
(рис. 4-7) Эксперименты с b_i, равными 5, 0, -2 и из N(0,1) соответственно

На графиках по оси абсцисс — текущий временной шаг, по оси ординат — количество зараженных агентов. Полностью зараженной будем считать сеть в момент с 99% зараженных от всех агентов. Это количество обозначено на графиках пунктирной линией. Последние агенты могут долго оставаться незараженными, если они сильно отделены от остальной сети.

Из графиков видно, что влияние агентов непосредственно определяет скорость каскада. Из-за диаметра сети полное заражение в худшем случае будет не быстрее 8 шагов.

4.2. Реактивное сопротивление

Пусть размерность вектор-параметра x равна 1. Разделим отношения между агентами (все ребра) случайно на доверчивые ($w_{ij} = (2)$) и скептические ($w_{ij} = (-2)$) и рассмотрим скорость распространения при разном соотношении таких отношений. Пусть $x = (1)$, влияние $b_i = 0$.

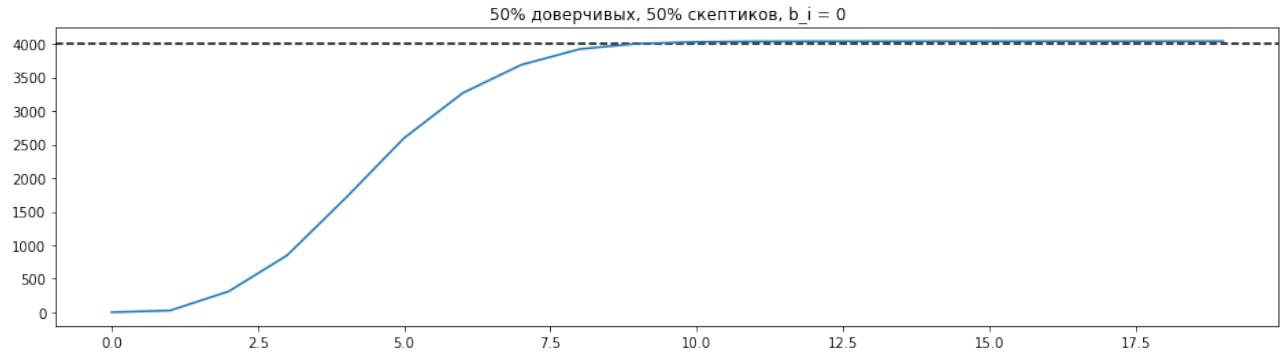


(рис. 8-11) Эксперименты с различными соотношениями доверчивых и скептиков

Заметно, что большее количество недоверия к основному параметру сообщения обеспечивает более долгое распространение каскада.

Можно подобрать оптимальный параметр x для наискорейшего распространения каскада при соотношении доверчивых и скептиков 50/50. Ищем его экспериментально на линейном пространстве на отрезке $(-1; 1)$.

Оптимальным параметром оказывается $x = 0$, что логично — в таком случае в силу равного соотношения скептиков и доверчивых в среднем получается обычная модель независимых каскадов с шансом заражения $p = 0.5$ для всех связей.



(рис. 12) Эксперимент с найденным оптимальным x

Подобные переборы можно делать для других соотношений и для других параметров x, используя метод Монте-Карло.

4.3. Параметр с двумя компонентами

Пусть размерность n вектор-параметра x равна 2. Подберем оптимальный параметр для социальной сети с нетривиальной демографией.

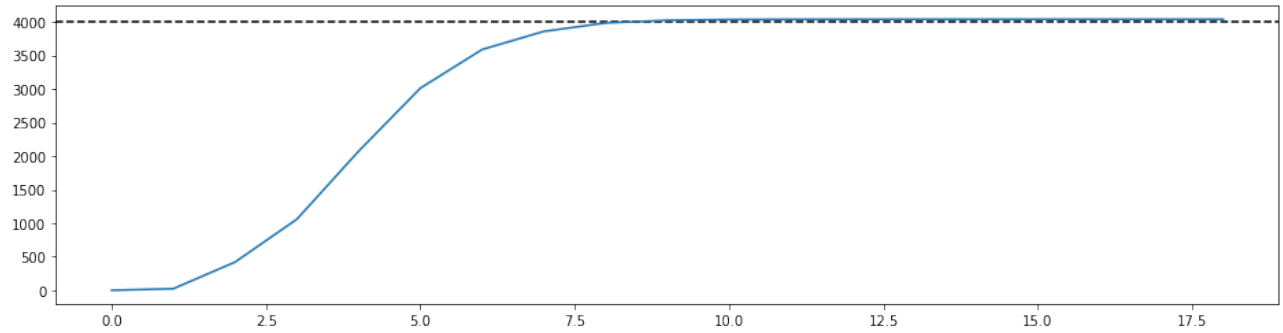
Рассмотрим гипотетическую страну, где $2/3$ — коренное населением, $1/3$ — мигранты. Положим, что коренное население поровну делится на либералов и консерваторов, а все мигранты — консерваторы.

Пусть формат нашего сообщения имеет две компоненты — политическая и национальная ориентированность. Политическая ориентированность, равная 1, обозначает полное соответствие консервативным ценностям, (-1) — либеральным. Национальная ориентированность, равная 1, обозначает упор сообщения на коренное население, (-1) — на мигрантов.

Коренные либералы обладают входящими ребрами с параметром $w_1 = (-2; 0)$, что обозначает, что им нравятся сообщения с либеральной повесткой, а национальная ориентированность им безразлична.

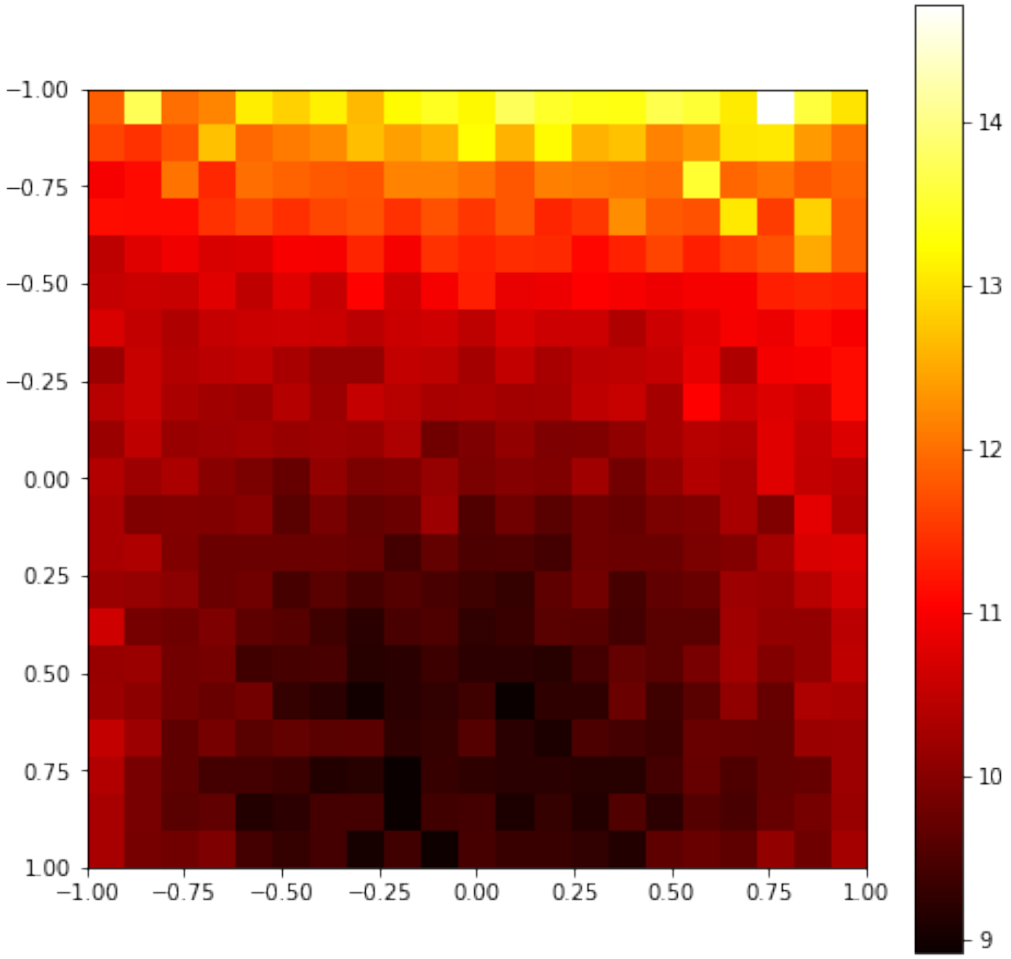
Ребра коренных консерваторов обладают параметром $w_2 = (2; 2)$ — им важна не только консервативная повестка, но и национальная ориентированность.

Мигранты обладают ребрами с параметром $w_3 = (2; -2)$ — им тоже важна консервативная повестка и национальная ориентированность.



(рис. 13) Пример: эксперимент на сети с $x = (0.5; 0.5)$

Составим карту зависимости времени, за которое заражается вся сеть, от компонент параметра $x = (x_1; x_2)$, которые мы ищем на отрезках $(-1; 1)$. Для этого были проведены расчеты для всех возможных комбинаций x_1 и x_2.

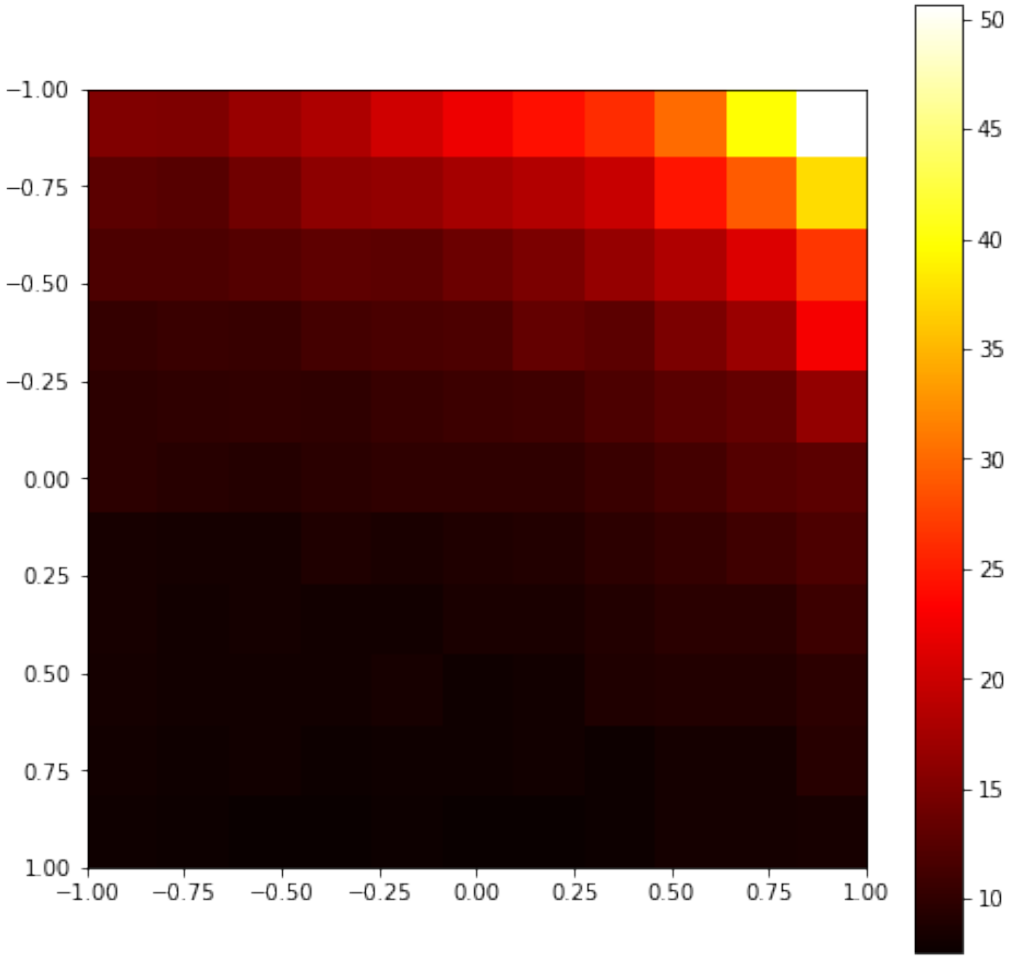


(рис. 14) Тепловая карта зависимости длительности каскада от формы сообщения

На карте по оси абсцисс — x_2, по оси ординат — x_1. Цветом обозначено, сколько шагов в среднем занимает распространение сообщения на 99% сети. Чем светлее цвет ячейки — тем больше времени занимает процесс распространения сообщения.

Из карты можно сделать вывод, что наиболее эффективные формы сообщения — с большим уклоном в консервативную повестку, без уклона в национальном ключе (на карте темное пятно). Сообщение с либеральной повесткой проходит дольше всего (светлая полоса сверху карты).

Если повысить долю мигрантов в обществе до 66%, то карта будет выглядеть иначе. Теперь скорость распространения информации в сети сильнее зависит от их мнения.



(рис. 15) Тепловая карта зависимости длительности каскада от формы сообщения в сети с большей долей мигрантов

Самая эффективная форма сообщения здесь — консервативная и направленная на мигрантов (нижний левый угол), наименее эффективная — либеральная и ориентированная на коренное население (верхний правый угол).

В примерах не учтено, что коренное население и мигранты в сети в реальности будут находиться в различных пространственных кластерах. Здесь они распределены равномерно. Также не учтены возможные разные влияния агентов — здесь они приравнены нулю. В реальности, эти дополнительные данные можно получить и использовать для поиска более оптимальной формы сообщения.

5. Вывод и перспективы работы

В работе была придумана и реализована модель информационных каскадов. На искусственной социальной сети показана действенность этой модели в отображении реальных процессов, а также представлен способ нахождения оптимальной формы сообщения для наискорейшего распространения информационного каскада. Модель обладает большими возможностями и перспективами для дальнейших экспериментов по поиску закономерностей распространения информации в обществе.

Результаты исследования можно применять на практике для рекламы продукта или распространения любой другой идеи. Рассмотрим конкретный случай на примере публичной страницы в социальной сети ВКонтакте. Владелец страницы имеет возможность видеть список всех ее подписчиков. Их нужно кластеризовать и рассмотреть присущие каждой группе интересы. Это можно сделать, рассмотрев, на какие другие страницы они подписаны, какие записи они пересылали на свою страницу. В конечном итоге, это делается специализированным программным обеспечением, реализующим методы машинного обучения. Если еще так же анализировать друзей подписчиков, то можно более точно понять, какие параметры для формы сообщения стоит иметь в виду при его создании, чтобы получить максимально эффективный информационный каскад.

Моделью не учтена возможность эволюции формата сообщения со временем. В реальности идеи все время меняются. Игра "сломанный телефон" — самый наглядный пример этого явления. Что-то можно по-своему понять, а что-то намеренно изменить при пересказе сообщения следующему слушателю. Чаще будут встречаться очень малые изменения формата с единичным шагом времени, но на больших отрезках времени они могут суммироваться в значимые изменения.

Внутренний диалог человека тоже можно считать за социальную связь (в этом случае ребро направлено из агента обратно в него). Это важно учесть при рассмотрении эволюции формата идеи.

Интересную тему для изучения представляет из себя противоборство двух или более сообщений, одновременно распростаняющихся в сети. Это, например, могут быть аналогичные продукты или повестки соперничающих политических партий. На основе предложенной модели передачи сообщений может получиться интересное взаимодействие эволюционирующих идей.

Графовые нейронные сети все чаще начинают использоваться в том числе на социальных сетях и, в частности, для моделирования передачи сообщения. Методы глубокого обучения можно использовать для получения наиболее точной оценки интересов и влияния агентов, до которых и через которых мы хотим донести сообщение, путем анализа их персональных страниц и активности в интернете.

6. Перечень использованных ресурсов

Список литературы

[1] Дмитрий Губанов, Дмитрий Новиков, Александр Чхарташвили. (2010). Социальные сети: модели информационного влияния, управления и противоборства.

[2] Andrea Montanari, Amin Saberi. (2010). The spread of innovations in social networks. Proceedings of the National Academy of Sciences of the United States of America. 107. 20196-201. 10.1073/pnas.1004098107.

[3] Duncan Watts, Peter Dodds. (2007). Influentials, Networks, and Public Opinion Formation. Journal of Consumer Research. 34. 441-458. 10.1086/518527.

[4] Gabriel Rossman and Jacob C. Fisher, "Network hubs cease to be influential in the presence of low levels of advertising", Proceedings of the National Academy of Sciences (USA) 118 (2021): e2013391118

[5] Ragia Ibrahim, Aboul Ella Hassanien, Hesham Hefny. (2018). Controlling Social Information Cascade A Survey: A Social Network Approach. 10.1201/9781315112626-9.

[6] Elmar Kiesling, Markus Günther, Christian Stummer, Lea Wakolbinger. (2012). Agent-based simulation of innovation diffusion: A review. Central European Journal of Operations Research. 20. 183-230. 10.1007/s10100-011-0210-y.

[7] https://ndg.asc.upenn.edu/experiments/creating-critical-mass/

[8] Marc Lelarge. (2011). Diffusion and cascading behavior in random networks. ACM SIGMETRICS Performance Evaluation Review. 39. 34. 10.1145/2160803.2160852.

[9] F. Altarelli, A. Braunstein, L. Dall'Asta, and R. Zecchina, "Large deviations of cascade processes on graphs"

[10] Duncan J. Watts, "A simple model of global cascades on random networks", Proceedings of the National Academy of Sciences (USA) 99 (2002): 5766--5771

[11] Moez Draief and Laurent Massoulié, "Epidemics and Rumors in Complex Networks"

[12] Онлайн-курс "Network Dynamics of Social Behavior", Damon Centola, University of Pennsylvania

[13] Л.Л. Делицын. "Разработка и применение количественных моделей распространения новых информационных технологий". Научно-техническая информация, серия "Организация и методика информационной работы", № 5, с. 24-32.