Методы анализа информационных потоков в сети Интернет

[1,2] *Аветисян А. А.<a.a.avetisyan@ispras.ru>*
[1,3] *Дробышевский М. Д.<drobyshevsky@ispras.ru>*
[1,2,4] *Турдаков Д. Ю.<turdakov@ispras.ru>*

[1]*Институт системного программирования им. В.П. Иванникова РАН, 109004, Россия, г. Москва, ул. А. Солженицына, д. 25*
[2]*Московский государственный университет имени М.В. Ломоносова, 119991, Россия, Москва, Ленинские горы, д. 1*
[3]*Московский физико-технический институт (государственный университет), 141701, Московская область, г. Долгопрудный, Институтский переулок, д.9*
[4] *НИУ "Высшая школа экономики", 101000, Россия, г. Москва, ул. Мясницкая, д. 20*

Аннотация. Распространение информации – это фундаментальный процесс, происходящий в сети Интернет. Ежедневно мы можем наблюдать публикацию различной информации и ее дальнейшее распространение через новостные агентства и сообщения обычных пользователей. И хотя сам процесс можно наблюдать явно, определить отдельные пути передачи очень сложно. Проникновение глобальной информационной среды во все сферы жизни человечества радикально меняет скорость и пути распространения информации. В этом обзоре мы исследуем модели распространения информационных потоков в сети Интернет, разделяя их на две группы: объяснительные, предполагающие наличие сети влияния между информационными узлами, и предсказательные, ставящие своей задачей изучение распространения отдельных частей информации. Несмотря на всю сложность, изучение глубинных свойств распространения информации необходимо для понимания общих процессов, происходящих в современном информационном обществе.

Ключевые слова: распространение информации, информационные каскады, пути распространения информации

DOI: 10.15514/ISPRAS-2018-30(6)-11

Для цитирования: Аветисян А.А., Дробышевский М.Д., Турдаков Д.Ю. Методы анализа информационных потоков в сети Интернет. Труды ИСП РАН, том 30, вып. 6, 2018 г., стр. 199-220. DOI: 10.15514/ISPRAS-2018-30(6)-11

1. Введение

Ежедневно в сети Интернет происходят различные процессы распространения информации: от публикации статей новостными агентствами до обмена сообщениями обычных пользователей. Учитывая огромное влияние распространения информации на жизнь современного общества, анализ и моделирование информационных потоков имеют важное значение для понимания происходящих процессов и изучения способов влияния на них. Высокоточные масштабные модели распространения и эволюции онлайновой информации необходимы как для анализа стратегических кампаний в информационном пространстве, так и для предоставления критически важной информации населению в ходе операций по оказанию помощи при бедствиях, а также могут потенциально способствовать решению других критических задач в онлайновой информационной области. При этом возникает ряд задач, связанных со сбором данных из разнородных источников, их анализом и построением математических моделей. Например, если появление новостей или сообщений можно наблюдать явно, то определить, кто является источником, в каждом случае часто затруднительно. Предположение о возможных сетях влияния можно сделать, собрав данные из блогов или новостных статей и проанализировав времена появления похожих сообщений в различных источниках [1].

Основные платформы для распространения информационных потоков включают: системы мгновенного обмена сообщениями, приложения для ведения блогов и микро-блогов, электронную почту, мобильные сети коммуникации и другие приложения, большинство которых используются на смартфонах [2]. При этом виды информации и способы ее распространения можно разделить на несколько категорий [2].

- Широковещание – кроме традиционного радио и телевидения, активно используются социальные медиа для маркетинговых кампаний и продвижений, рассылки предупреждающих сообщений, привлечения пользователей и так далее.
- Обмен контентом в онлайновых соцсетях (sharing), доступный каждому пользователю: посты, репосты, ретвиты и другие, – согласно исследованию на основе Фейсбука [3], главными мотивами становятся «получение удовольствия» и «раскрытие информации, необходимой/полезной для извлечения выгоды из Фейсбук».
- Краудсорсинг – предполагает привлечение широких групп людей для решения задач путем обмена идеями и контентом, особенно внутри онлайн-сообществ.
- С развитием соцсетей активно используется вирусный маркетинг (viral marketing), целью которого является продвижение товаров и услуг среди пользователей путем распространения рекламы в видео, изображениях, блогах, текстовых сообщениях при условии передачи их между людьми.

В результате упомянутых процессов происходит формирование мнений (opinion formation) как у отдельных индивидов, так и у целых сообществ относительно новостей, политических событий или новых продуктов [4]. Аналогичным образом происходит распространение инноваций (innovation diffusion), например, в результате перенятия людьми идей от друзей в социальной сети [5].

- Отдельно стоит выделить распространение вредоносной информации (malicious spreading), такой как компьютерные вирусы, вредоносная реклама, слухи и сплетни. Благодаря высокой скорости передачи через Интернет сети, массовое заражение компьютерными вирусами представляют серьезную опасность [6], а распространение дезинформации может иметь неприятные последствия для общества [7].

Данная статья посвящена обзору методов анализа информационных потоков в сети Интернет, разработанных за последние годы. В разд. 2 приведены важные определения и основные свойства информационных потоков. Далее, в разд. 3 описаны методы и модели, в соответствии с разными аспектами задачи. В разд. 4 говорится о приложениях этих моделей, и затем следует заключение.

2. Основные свойства информационных потоков

Центральным понятием при изучении информационных потоков является **информационный каскад**. В сети появляется новая единица информации – сообщение, вброс, новость или вирус – которая на некоторый ограниченный период времени вызывает повышенное внимание. Тема последовательно появляется в различных узлах сети, захватывая таким образом все больший круг узлов и образуя каскад. Обычно отдельный каскад моделируется деревом, корнем в котором становится пользователь, первым опубликовавший сообщение, а информация распространяется от родителя к потомкам. Однако в других случаях могут быть несколько входящих ребер, например, когда один пользователь агрегирует множество рекомендаций.

2.1 Временн*ые* динамические свойства

Наиболее общей характеристикой информационных потоков служит последовательность их появления, бурного роста и спада (the rise and fall pattern), притом, что разные темы непрерывно сменяют друг друга. Это проиллюстрировано на рис. 1, где приведены 50 наиболее масштабных новостных тем за период с 1 августа по 31 октября 2008 года, собранных с медийных сайтов Google News (20,000 сайтов) и 1.6 млн блогов, форумов и других сайтов. В первых попытках описания динамических паттернов используется пуассоновский процесс, а функции роста и падения приближаются с помощью степенного закона с экспонентой от -0.1 до -2.5 [9]. При анализе хэштегов Твиттера было выявлено несколько классов временных паттернов, выражающих активность до и во время пика, во время и после

пика, симметрично по обе стороны от пика и на следующий день после пика [10]. Паттерн взлета и падения также может быть описан стохастическими дифференциальными уравнениями при ряде допущений и упрощений [11].

Однако динамические свойства информационных потоков пока остаются мало изученными, что сильно препятствует возможности предсказания каскадов.

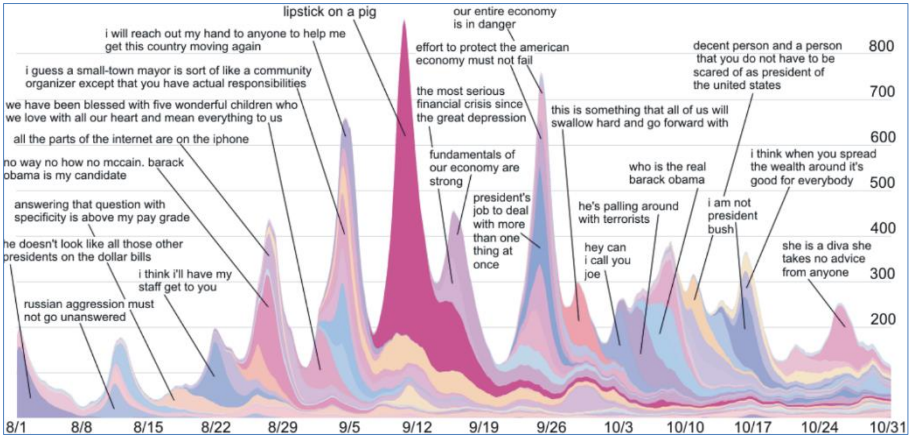

Рис. 1. 50 наиболее масштабных потоков в новостных циклах за период 1 августа – 31 октября 2008 года [8]. Каждый поток состоит из всех новостных заметок и постов в блогах, содержащих в своем тексте определенные фразы (некоторые приведены на рисунке)

Fig. 1. 50 most large-scale threads in news cycles for the period August 1 – October 31, 2008 [8]. Each thread consists of all news notes and blog posts that contain certain phrases in their text (some are shown in the figure)

Большое разнообразие паттернов обусловлено сложным устройством социальных систем. С одной стороны, это тип контента: было показано наличие существенных различий в характере распространения хэштегов в Твиттере в зависимости от их темы [12], а также в динамике каскадов в Фейсбуке в зависимости от категории распространяемой новости [13].

Вторым фактором является среда распространения, наиболее важную роль здесь играет структура лежащей в основе сети. В целом, пользователи с большим количеством друзей/подписчиков имеют большее влияние как на скорость распространения, так и на популярность передаваемой информации [14].

Также важную роль играют и другие свойства графа: распределение степеней, эффект ``малого мира'', структура эго-сетей, сила связей, идеологическая гомофилия при образовании связей. Например, было обнаружено, что, хотя сильные связи оказываются влиятельными локально, более многочисленные слабые связи ответственны за передачу новой информации [15].

Еще бо́льшую трудность несет тот факт, что топология сетей непостоянна во времени, ребра могут менять свою активность в рассматриваемых процессах, а

также само распространение информации оказывает влияние на эволюцию связей в сети.

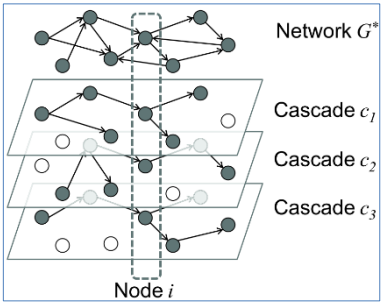

Рис. 2. Сеть распространения информации, полученная объединением путей передачи индивидуальных каскадов [1]
Fig. 2. Information diffusion network obtained by combining the paths of transmission of individual cascades

2.2 Структурные свойства

Имея данные о множестве информационных каскадов для одних и тех же участников, можно построить общую сеть распространения (diffusion network), объединив индивидуальные пути передачи из отдельных каскадов (рис. 2). В таких сетях экспериментально исследуются частота встречаемости каскадов определенной структуры и распределение их размеров. Несмотря на существование «вирусных» каскадов, достигающих огромных размеров, при анализе различных доменов – от коммуникационных платформ до игровых сетей и микроблогов – было обнаружено, что подавляющая доля (73-95%) каскадов состоит лишь из одного узла, то есть распространения не происходит [16]. Распределения и размера каскада, и его глубины (высоты дерево), измеренные для разных платформ (Facebook, Digg, Sina Weibo, Slashdot и других), показывают перекос в сторону каскадов малого размера и малой глубины. Обычно полагают, что обе зависимости приблизительно описывается степенным законом. Таким образом, большие каскады представлены относительно редко.

Для характеристики структуры каскада была предложена структурная виральность (structural virality, известная также как Wiener [17]):

$$v = \frac{1}{n(n-1)} \sum_i \sum_j d_{ij}$$

выражающая среднее расстояние d_{ij} между всеми парами вершин каскада. Низкое значение виральности говорит о распространении через узлы-хабы (больше в ширину), а высокое значение – о наличии длинных путей (больше в

глубину). Для многих систем были показаны низкие значения виральности вместе с низкой корреляцией значения виральности и размера сети распространения.

В [18] предлагается индекс потенциала виральности (Virality Potential Index), учитывающий для каждого узла m сети долю его потомков $Follow(m)$, которые участвуют в каскаде: $Infected(m)$. В предположении, что вероятность передачи следующим потомкам одинакова и равна «потенциалу» передачи $\frac{Infected(m)}{Follow(m)}$, индекс определен как минус логарифм от вероятности для всего дерева:

$$VPI(m) = \sum_{m} -|Infected(m)| \, log \frac{|Infected(m)|}{|Follow(m)|}$$

Индекс минимален, когда эта доля равна нулю (никто ничего не передал) или единице (нет потенциала передать больше). Индекс для каскада считается как средний индекс по всем его элементам m.

Кроме распространения информации в сети от одного узла к другому, имеет место приток из внешних источников, таких как масс-медиа, новостные сайты и телевидение. Например, анализ упоминаний URL в Твиттере показал, что на P2P взаимодействия приходится около 71% объема информации [19], а на популярность хэштегов внешние источники оказывают даже большее влияние [10].

2.3 Свойства поведения пользователей

Разнообразное поведение пользователей в социальной сети также будет генерировать различные шаблоны распространения информации. Влияние поведения пользователей может зависеть от разных факторов. Пользователи с одинаковыми предпочтениями с большей вероятностью будут взаимодействовать друг с другом, что существенно влияет на распространение информации. Анализ временны*ы*х признаков в системе мобильной связи выявляет, что социальное взаимодействие более часто встречается у лиц с аналогичными признаками, такими как пол и возраст. В работе [20] авторы наблюдают идеологическую связь в сетях друзей, основываясь на данных Facebook, где консерваторы/либералы с большей вероятностью дружатся с человеком с той же политической принадлежностью. Помимо влияния на структуру социальной сети, интерес пользователей или предпочтение также будет напрямую влиять на распространение информации. Как правило, пользователи хотели бы оценивать и передавать информацию, которая согласуется с их интересами, а разные предпочтения обычно интерпретируются как вероятность распространения. И факторы среды распространения, такие как информационный контент, социальное влияние или даже эмоции, будут существенно влиять на предпочтения пользователей.

Одним из факторов влияния поведения пользователей является динамическое событие с временными увеличениями интереса. Обычно поведение индивидов демонстрирует временный всплеск, где часто появляющиеся события наблюдаются в течение очень коротких периодов, за которыми следуют длительные периоды бездействия. Васкес и др. [21] анализируют распространения почтовых червей среди пользователей электронной почты. Предполагается, что временной интервал между двумя последовательными сообщениями, отправленными одним и тем же пользователем почты, является непуассоновским процессом. Эта гипотеза согласуется с наблюдаемыми процессами распространения вирусов, которые показывают время угасания, близкое к году, в отличие от однодневного угасания, предсказанного моделями, основанными на пуассоновском процессе.

2.4 Другие закономерности

Определение силы влияния для каждого из узлов дает увидеть некоторые закономерности в распространении информации. Например, после определения функции влияния различных веб-сайтов обнаружено, что они в значительной степени зависят от типа веб-сайта и темы информации. Например, если передача коротких текстовых фраз, связанных с новостями, сильно зависит от влияния нескольких крупных медиа-сайтов, то передача хэштегов Twitter регулируется гораздо большим набором активных пользователей, каждый из которых имеет относительно меньшее влияние. Также отмечено, что пользователи с наибольшим числом подписчиков не являются наиболее влиятельными в распространении хэштегов [22].

Предсказание большого каскада затруднено из-за множества факторов, влияющих на процесс диффузии информации. Проблема предсказания большого каскада рассматривается из-за редкого появления больших каскадов во многих системах и непредсказуемости в социальных системах. В работе [23] пришли к выводу, что никакая мера качества не может точно предсказать величину распространения, с помощью искусственного эксперимента «музыкального рынка», который показал, что песни с одинаковыми начальными условиями могут достичь очень разных уровней популярности.

В [24] авторы показали возможность предсказания того, что каскад достигнет наибольшей величины, с высокой точностью с учетом временных и структурных признаков начального процесса расширения. Более точное предсказание можно было бы дать, имея больше информации о каскаде, например, о большей части пути в каскаде распространения. Но до какой степени можно предсказать распространение информации и какие функции являются наиболее значимыми для прогнозирования информационных каскадов, все еще неясно. Хотя было бы очень трудно идентифицировать внутренние свойства для предсказания информационного каскада, некоторые закономерности получаются на основе эмпирического анализа. Например, важность индивидуумов, которые участвуют в каскаде на начальных этапах, и

ширина (а не глубина) каскадной структуры хорошо коррелирует с окончательным размером каскадов.

3. Метрики

В данном разделе описываются данные, используемые в изучении методов анализа информационных потоков, и различные модели распространения информации

3.1 Датасеты

В основной части исследований моделей предсказания используются данные, по которым легко отследить пути передачи информации, поэтому в большинстве случаев данными являются наборы твитов из Twitter или Sina Weibo или посты Facebook. Большинство данных находится в закрытом доступе. Для каждого твита или поста определен автор и первоначальный источник информации, а передачей информации от одного пользователя другому считается ретвит (репост) информации. В исследованиях также используют информационные каскады, собранные из новостных блогов и СМИ. Пути распространения информации от одного узла к другому создаются с помощью инструмента Memetracker [25], который строит карты ежедневного цикла новостей, анализируя около 900 000 новостных сообщений и сообщений в блоге в день из 1 миллиона онлайн-источников, а также гиперссылки между новостными сайтами, ссылающимися на одну и ту же информацию.

3.2 Объяснительные модели

В этом подразделе описываются модели, в которых предлагаются различные механизмы, объясняющие наблюдаемые информационные потоки. Пороговые (п. 3.2.1), каскадные модели (п. 3.2.2) и модели эпидемий (п. 3.2.3) основываются на существующей сети связей между узлами. При отсутствии такой сети в явном виде, другой тип моделей (п. 3.2.4) предполагает наличие скрытых отношений между узлами.

3.2.1 Пороговые модели

Основная идея заключается в том, что люди в сети принимают решения, основанные на действиях своих соседей. Такая стратегия принятия решений подразумевает, что пороговая модель (threshold model) содержит память об истории распространения. Описывается она следующим образом: на первом шаге произвольным образом выбирается небольшая часть людей, которая получила информацию с самого начала. В течение каждого временного шага для пользователя i в неактивном состоянии определяется ϕ_i как пороговое значение i. Состояние i изменится на активное, если доля его соседей в активном состоянии равна или больше, чем ϕ_i. Все узлы в активном

состоянии остаются без изменений. Процесс диффузии заканчивается, когда число пользователей в активном состоянии не меняется.

Чентола и др. [26] применяют пороговую модель для различных социальных сетей. В этой статье пришли к выводу, что, в отличие от результатов моделей независимого взаимодействия, где влияние пары узлов между собой не связано с влиянием других, случайные передачи информации между удаленными узлами уменьшают способность сетей распространять информацию.

Одно из применений пороговой модели заключается в изучении влияния структуры сообщества на распространение информации. Так, используя линейную пороговую модель, Нематзаде и др. [27] раскрыли роль модульной структуры в распространении информации, которая указывает на то, что влиятельные сообщества могут усилить локальное распространение, а слабые сообщества могут улучшить глобальное распространение. Более того, найдена оптимальная модульная структура, которая может способствовать как локальному, так и глобальному распространению.

В различных статьях предлагаются обобщения пороговой модели. Например, Доддс и Уоттс [28] формулируют модель, которая учитывает память о прошлых воздействиях на узлы сети. Модель также включает в себя Independent interaction model (SIS и SIR из п. 3.2.3) в качестве частного случая. Брамитт и др. [29] предполагают, что узел становится активным, если доля активных соседей в любом слое превышает пороговое значение. Авторы [30] модифицируют линейную пороговую модель, предполагая, что на каждого пользователя влияют его соседи в течение некоторого периода времени.

3.2.2 Каскадные модели

Каскадная модель (cascade model) основывается на двух гипотезах. Во-первых, попарные влияния узлов друг на друга независимы, во-вторых, любой активный узел i имеет только один шанс передать информацию своему соседу j, независимо от результата на последующих шагах узел i не будет влиять на j. На первом шаге, как и в пороговой модели выбирается некоторый набор узлов, который активируют. На новом шаге k каждый узел u, получивший информацию на предыдущем шаге, пытаются активировать всех неактивированных соседей j с вероятностью p_{ij}. Процесс останавливается, когда новые пользователи, получившие информацию, не появляются.

В реальных сетях возникает проблема предсказания вероятности всех ребер распространения. Одно из решений этой проблемы предложено в работе Диккенса и др. [31], в которой рассматриваются масштабируемые методы как для обучения информационных потоков на основе моделей независимых каскадов, так и для предсказания нового потока. Определяются два типа

данных: данные, для которых известны пути предыдущих потоков, и данные, для которых известны только конечные точки распространения.

При рассмотрении задачи распространения информации формулируется две проблемы разного типа: проблема максимизации влияния и минимизации «заражения». В первом случае стоит задача максимизировать конечный размер информационного каскада, во втором требуется минимизировать его влияние на сеть.

Предположим, что A – множество изначально активных узлов, ожидаемое число активных узлов в конечном состоянии обозначается как $\sigma(A)$. Задача максимизации влияния требует найти множество узлов S размера k такое, что $\sigma(A)$ максимально. Кемп и др. [32] изучают проблему максимизации влияния как дискретную задачу оптимизации, основанная на двух моделях: независимой каскадной модели и линейной пороговой модели, для которых применяется жадный алгоритм поиска восхождением. Для жадного алгоритма требуется большое количество вычислений, так как предельный выигрыш для $\sigma(A)$ должен вычисляться много раз при разных наборах начальных узлов. Кимура и др. [33] предложили эффективный метод оценки $\sigma(A)$, основанный на теории перколяции и теории графов, который применяется к приближенному решению задачи максимизации по жадному алгоритму. Эксперименты, проведенные на крупномасштабных сетях реального мира, показали, что этот метод может сократить вычислительную сложность.

Для решения проблемы «заражения» в ранних работах [34] из сети удалялись узлы. Кимура и др. [35] предложили новый метод блокировки ограниченного числа звеньев в сети для уменьшения размера каскада. В этой работе определяется степень «заражения» $C(G)$ на графе $G = (V, E)$, которая является средним от степеней влияния всех узлов v в G, где степень влияния узла v в G, определяется как ожидаемое число активных узлов в конце процесса распространения для начального активного узла v. Поскольку решить данную проблему в больших сетях вычислительно трудно, предлагается использовать жадный алгоритм для получения приближенного решения этой проблемы.

3.2.3 Модели эпидемий

В моделях эпидемии (epidemic models) люди разбиваются на различные группы, и каждой группе назначается какое-то состояние. Три состояния являются наиболее часто используются в процессе диффузии:

- S – восприимчивое состояние (susceptible), в котором узел еще не заражен;
- I – зараженное состояние (infectious), представляющее уже активированных пользователей, которые могут передать вирус другим;
- R – восстановленное состояние (recovered), в котором узел заражен, но не является активным и больше не будет передавать вирус другим.

Различные комбинации этих состояний могут приводить к различным моделям, таким как SI, SIS и SIR.

- Простейший случай модели эпидемии – модель SI, в которой рассматриваются два состояния: S и I. Эта модель похожа на пороговую модель и каскадную модель, поскольку, когда узел становится активированным, он остается в этом состоянии навсегда.
- Модель SIS используется для характеристики эпидемий, имеющих временный иммунитет. Это означает, что зараженные люди снова становятся восприимчивыми.
- Модель SIR вводится для объяснения эпидемий с постоянным иммунитетом среди населения. В отличие от модели SI и SIS, в этой модели приводится восстановленное состояние R. Узлы S будут активированы пользователями состояния I с вероятностью β, а узлы I будут восстанавливаться до состояния R с вероятностью восстановления γ.

Гомес-Родригес и др. [1] также предполагали, что их метод применим не только для моделирования информационных потоков, но и для предсказания распространения вирусов. Однако более поздние работы [36] показали, что процесс распространения информации в социальных медиа отличается от процессов распространения вирусов. Так, например, люди с высокой степенью связности (большим количеством подписок) с меньшей вероятностью передают информацию дальше. Как следствие, существующие модели, основанные на моделировании процесса заражения, не могут в полной мере описать процесс распространения информации.

3.2.4 Модели на основе скрытых связей

Для объяснения распространения информации предполагается наличие некоторых независимых скрытых отношений между пользователями, которые необходимо изучить. Так, Гомес-Родригес и др. [1] делают предположение о том, что существует некоторая базовая статическая сеть, по которой распространяется информация. Можно наблюдать, когда поток достигает узла сети, но неизвестно, откуда информация к нему попала. Анализируется набор независимых информационных каскадов, для которых известен набор пользователей, получивших информацию и время ее получения. Вычисляется вероятность наблюдения набора каскадов в выбранной скрытой сети, и задачей является построение наиболее вероятной скрытой сети, объясняющей распространение данных каскадов. Вводится функция правдоподобия, задача поиска максимума которой эквивалентна нахождению наиболее вероятного графа. Предлагается итеративный жадный алгоритм NetInf максимизации функции правдоподобия, где на первом шаге выбирается пустой граф на вершинах-пользователях сети, а на каждом новом шаге добавляется ребро с наибольшим вкладом в функцию правдоподобия.

В отличие от Netinf, фиксирующего скорость распространения информации между узлами, в работе Гомеса-Родригеса и др. [37] предлагается модель Netrate, которая дополняет Netinf тем, что допускает передачу информации по ребрам с разной скоростью. Вероятность узла, активирующего другой в данный момент времени, моделируется функцией плотности вероятности в зависимости от времени активации и скорости передачи между двумя узлами. Алгоритм Netrate, помимо построения графа распространения, вычисляет скорости передачи информации между узлами.

Недостаток предыдущих моделей в том, что их сеть влияния пользователей остается статической с течением времени. По этой причине Гомес-Родригес и др. [38] расширяют модель Netrate и предлагают алгоритм InfoPath, в котором используются стохастические градиенты построения графа влияния в каждый момент времени (раз в день).

Буриго и др. [39] предлагают модель для предсказания распространения информации и влияния узлов друг на друга. Ее отличие состоит в том, что для каждого пользователя определяется вектор в пространстве. Вероятности влияния выводятся из относительных положений векторов соответствующих пользователей в этом пространстве. Вектора пользователей вычисляются на основе данных каскадов.

3.3 Предсказательные модели

Другой класс моделей нацелен на предсказание того, как определенный информационный поток будет распространяться в данной сети на основе свойств предыдущих диффузий. Например, предсказание количества репостов для данного сообщения или дальнейшей динамики диффузии. Среди моделей для предсказания распространения информации можно выделить два подхода: модели на основе ролей пользователей (п. 3.3.2) и предсказание на основе признаков (п. 3.3.1).

3.3.1 Модели на основе признаков

Модели на основе признаков решают задачу предсказания распространения информационного каскада на основе некоторых свойств диффузии. Авторы, описывающие модели на основе признаков, используют данные, в которых легко отследить передачу информации от одного пользователя к другому. Такими, например, являются Twitter, Facebook, Sina-Weibo. В дальнейшем в этом разделе прием информации одним пользователем от другого будем называть репостом.

Одной из задач изучения распространения информации, является предсказание популярности новости. Для обучения моделей используются свойства новости и атрибуты пользователя и применяются стандартные бинарные классификаторы [40-41] или методы глубокого обучения [42-43]. Чэн и др. [24] предлагают следующую формулировку: для уже

распространяющегося информационного каскада предсказать, удвоится ли его размер. Результаты показывают, что точность такого предсказания растет с размером каскада; каскады, больше растущие в ширину, чем в глубину, с более высокой вероятностью достигнут больших размеров.

Шульман и др. [44] пытаются для потока новостей предсказать, станет ли он более популярен, чем определенный процент других новостей, учитывая набор новостей и данных об истории их распространения. Авторы статьи приходят к выводу, что временные признаки потоков наиболее важные в предсказании популярности распространения, при снижении влияния временных признаков точность значительно уменьшается.

В [45-46] представлен частный случай предсказания популярности: предсказание, будет ли у новости хотя бы один репост. В работе Петровича и др. [45] используется passive-aggressive алгоритм. Помимо глобальной модели обучаются 24 локальные модели для каждого часового периода в сутках, что позволило улучшить качество. Чжан и др. [46] для предсказания репоста предложил сверточную нейронную сеть, в которой объединены признаки пользователя, автора, пользовательские интересы, содержание твита и сходство между интересами пользователя и твитом. Цзян и др. [47] предлагают модель предсказания репостов, основанную на вероятностном методе матричной факторизации, путем интеграции наблюдаемых данных, социального влияния пользователей друг на друга и семантики сообщений.

3.3.2 Модели, учитывающие внешние источники информации

Один из способов улучшить предсказание распространения информации – понять, какую роль играет каждый пользователь в сети распространения. Ян [48] предлагает социальную модель распространения информации RAIN для изучения социальных ролей пользователей и моделирования распространения информации одновременно. RAIN определяет распределение социальных ролей каждого пользователя в соответствии с его структурными атрибутами и его поведением в процессе диффузии. Пользователи разделяются на три социальные группы: лидеры мнений (opinion leaders), пользователи, соединяющие разные сообщества (structural hole spanners), и обычные пользователи (ordinary users). Вводятся параметры для каждой роли ρ_r и λ_r как вероятность того, что пользователи, играющие роль *r*, активируют другого пользователя успешно и соответственно вызовут задержку распространения в течении одной временной отметки. Модель состоит из двух основных частей: сначала для каждого пользователя предсказывается одна из трех возможных ролей, после чего используется функция диффузии (например, пороговая или каскадная функцию), параметризованная ρ_r и λ_r, чтобы определить, станет ли пользователь активным.

Чубдар и др. [49] также предложили модель разбиения по ролям. Пользователей делят на пять ролей влияния, основываясь на различных

структурных свойствах (рейтинге влиянию на другие узлы, рейтинге принятия информации от других узлов, возрасте пользователя и т.д). Модель отличается от [48] тем, что роли пользователей непостоянны. Предложена динамическая модель кластеризации, которая объединяет свойства пользователя и его поведение в течение времени.

Если предыдущие модели предлагают классификацию пользователей по ролям, то Янг и Лесковец [22] рассматривают модель, в которой у каждого пользователя нет конкретной роли, но есть уровень влияния. По сети распространяются независимые каскады, и формулируется линейная модель влияния LIM (linear influence model), исходя из предположения, что количество узлов, вновь получивших информацию, зависит от того, у каких узлов эта информация была в прошлом. Затем число вновь активированных узлов моделируется как функция от времени, когда другие узлы получали информацию в прошлом. В этой модели каждый узел имеет связанную с ним функцию влияния. Тогда число вновь активированных узлов в момент времени *t* является функцией влияния узлов, зараженных до времени *t*.

4. Приложения

Изучение проблемы распространения информационных потоков в сети применимо во многих областях деятельности. Вирусный маркетинг является одним из наиболее важных приложений. Маркетинг из уст в уста – это новая форма маркетинга продукта, которая позволяет максимально использовать сетевые эффекты (например, через различные социальные сети) для повышения осведомленности о конкретном продукте и достижения рекламных целей. Он происходит в различных формах, включая изображения, видео, электронные письма, текстовые сообщения, твиты, игры и блоги. Компания может предоставить определенный продукт выбранному числу влиятельных лиц бесплатно, надеясь, что они порекомендуют продукт другим, если они будут им удовлетворены. Продавцы также могут предлагать скидки, основанные на влиятельности людей, в результате чего доход может даже быть увеличен. Целевая иммунизация – еще один пример изучения информационных потоков в сети, когда для новости требуется иммунизировать небольшое подмножество влиятельных узлов, чтобы уменьшить распространение передаваемой информации.

Распространение информации через социальные сети оказалось мощным инструментом во многих ситуациях, например, исследовались влияние Twitter на выборы президента [50] и Facebook в арабской весне 2010 года [51]. Другим важным исследованием в этой области являются эпидемии и распространение вирусов, которые привлекли многих ученых в области экологии и биологии. Идентификация узлов с наиболее важной ролью в распространении информации будет иметь множество потенциальных приложений в различных областях. Например, путем определения наиболее влиятельных распространителей в преступных группах, соответствующий

разведывательный орган может лучше контролировать преступность и осуществлять превентивные меры.

Диффузия инноваций – еще одна тема, которая часто изучается в сети. Новые идеи, технологии и способы выполнения действий могут быстро распространяться по сети. Было показано, что люди склонны принимать технологию (или продукт) с большей вероятностью, основываясь на мнении своих друзей и соседей (которые определяются на основе их связей в сети), которым они доверяют.

5. Заключение

В обзоре были рассмотрены методы анализа информационных потоков. Были выделены две основные группы рассматриваемых задач в процессе распространения информации.

Первая группа задач основана на предположении о существовании некоторой сети связей между узлами, по которой переходит информация. Пороговые и каскадные модели для существующих сетей пытаются определить, как будет распространяться каскад, основываясь на гипотезах получения информации, зависящих от доли активированных соседей и вероятности одного узла передать информацию другому. Модель, основанная на скрытых связях, предполагает наличие скрытой сети и пытается по информационным каскадам сети ее построить.

Во задачах второй группы не предполагается наличие скрытой сети, ее задачей является предсказание распространения определенного информационного потока на основе его признаков диффузии. Предлагается модель на основе признаков, которая, анализируя свойства информации и атрибуты пользователей, получивших ее, для уже распространяющегося информационного каскада предсказывает его популярность. Также, чтобы улучшить предсказание распространения, предложена модель, которая изучает влияние пользователя на другие узлы и предполагает, какую роль играет пользователь в распространении.

Несмотря на обилие работ в области, все они обладают рядом ограничений и недостатков. Большинство работ основаны на явно наблюдаемых механизмах социальных сетей (репосты), применяются только к одной социальной сети: Twitter, Sina Weibo или Facebook. В силу того, что большая часть данных, используемых в исследованиях, находится в закрытом доступе, предсказательные модели тестируются на разных данных, включая те, которые решают одну и ту же задачу. Это создает проблему сравнения результатов. Также не используются методы глубокого анализа текстов ввиду сложности разработки инструментов анализа текстов, способных обрабатывать большие объемы данных.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №18-07-01059.

Список литературы

[1] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 1019-1028.

[2] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, and Y.-C. Zhang. Dynamics of information diffusion and its applications on complex networks. *Physics Reports*, vol. 651, 2016, pp. 1-34.

[3] Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In Proc. of the International workshop on privacy enhancing technologies. Springer, 2006, pp. 36-58.

[4] S. Fournier and J. Avery. The uninvited brand. Business horizons, vol. 54, no. 3, pp. 193-207, 2011.

[5] G.E. Kreindler and H.P. Young. Rapid innovation diffusion in social networks. In Proceedings of the National Academy of Sciences, vol. 111, supplement 3, 2014, pp. 10 881-10 888.

[6] T. Holz, M. Steiner, F. Dahl et al. Measurements and mitigation of peer-to-peer-based botnets: A case study on storm worm. In Proc. of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, 2008.

[7] B. Doerr, M. Fouz, and T. Friedrich. Why rumors spread so quickly in social networks. Communications of the ACM, vol. 55, no. 6, 2012, pp. 70-75.

[8] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 497-506.

[9] Y. Sano, K. Yamada, H. Watanabe, H. Takayasu, and M. Takayasu. Empirical analysis of collective human behavior for extraordinary events in the Blogosphere. Physical Review E, vol. 87, no. 1, 2013.

[10] J. Lehmann, B. Goncalves, J.J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In Proc. of the 21st international conference on World Wide Web, 2012, pp. 251–260.

[11] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In Proc. of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 6–14.

[12] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In Proc. of the 20th international conference on World wide web, 2011, pp. 695–704.

[13] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H.E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. In Proceedings of the National Academy of Sciences, vol. 113, no. 3, 2016, pp. 554–559.

[14] S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. M. Tripathy, and S. Triukose. Spatio-temporal and events based analysis of topic popularity in twitter. In Proc. of the 22nd ACM international conference on Information & Knowledge Management, 2013,

pp. 219–228.
[15] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In Proc. of the 21st international conference on World Wide Web, 2012, pp. 519–528.
[16] S. Goel, D.J. Watts, and D.G. Goldstein. The structure of online diffusion networks. In Proc. of the 13th ACM conference on electronic commerce, 2012, pp. 623–638.
[17] S. Goel, A. Anderson, J. Hofman, and D. J. Watts. The structural virality of online diffusion. Management Science, vol. 62, no. 1, 2015, pp. 180–196.
[18] S. Krishnan, P. Butler, R. Tandon, J. Leskovec, and N. Ramakrishnan. Seeing the forest for the trees: new approaches to forecasting cascades. In Proc. of the 8th ACM Conference on Web Science, 2016, pp. 249–258.
[19] S.A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 33–41.
[20] E. Bakshy, S. Messing, and L.A. Adamic. Exposure to ideologically diverse news and opinion on facebook. Science, vol. 348, no. 6239, 2015, pp. 1130–1132.
[21] A. Vazquez, B. Racz, A. Lukacs, and A.-L. Barabasi. Impact of non-poissonian activity patterns on spreading processes. Physical review letters, vol. 98, no. 15, 2007.
[22] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In Proc. of the 2010 IEEE 10th International Conference on Data Mining (ICDM), 2010, pp. 599–608.
[23] M.J. Salganik, P.S. Dodds, and D.J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. Science, vol. 311, no. 5762, 2006, pp. 854–856.
[24] J. Cheng, L.Adamic, P.A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In Proc. of the 23rd international conference on World wide web, 2014, pp. 925–936.
[25] MemeTracker data. Режим доступа: http://www.memetracker.org/data.html, дата обращения 20.11.2018.
[26] D. Centola, V. M. Eguıluz, and M.W. Macy. Cascade dynamics of complex propagation. Physica A: Statistical Mechanics and its Applications, vol. 374, no. 1, 2007, pp. 449–456.
[27] A. Nematzadeh, E. Ferrara, A. Flammini, and Y.-Y. Ahn. Optimal network modularity for information diffusion. Physical review letters, vol. 113, 2014, no. 8.
[28] P.S. Dodds and D.J. Watts. A generalized model of social and biological contagion. Journal of theoretical biology, vol. 232, no. 4, 2005, pp. 587–604.
[29] C.D. Brummitt, K.-M. Lee, and K.-I. Goh. Multiplexity-facilitated cascades in networks. Physical Review E, vol. 85, no. 4, 2012.
[30] F. Karimi and P. Holme. Threshold model of cascades in empirical temporal networks. Physica A: Statistical Mechanics and its Applications, vol. 392, no. 16, 2013, pp. 3476–3483.
[31] L. Dickens, I. Molloy, J. Lobo, P.-C. Cheng, and A. Russo. Learning stochastic models of information flow. In Proc. of the 2012 IEEE 28th international conference on data engineering, 2012, pp. 570–581.
[32] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 137–146.
[33] M. Kimura, K. Saito, and R. Nakano. Extracting influential nodes for information

diffusion on a social network. In Proc. of the Twenty-Second AAAI conference on Artificial intelligence, vol. 2, 2007, pp. 1371–1376.

[34] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. Computer networks, vol. 33, no. 1-6, 2000, pp. 309–320.

[35] M. Kimura, K. Saito, and H. Motoda. Minimizing the spread of contamination by blocking links in a network. in Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, 2008, pp. 1175–1180.

[36] K. Lerman. Information is not a virus, and other consequences of human cognitive limits. Future Internet, vol. 8, no. 2, 2016.

[37] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. arXiv preprint arXiv:1105.0697, 2011.

[38] M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf. Structure and dynamics of information pathways in online media. In Proc. of the sixth ACM international conference on Web search and data mining, 2013, pp. 23–32.

[39] S. Bourigault, S. Lamprier, and P. Gallinari. Representation learning for information diffusion through social networks: an embedded cascade model. In Proc. of the Ninth ACM International Conference on Web Search and Data Mining, 2016, pp. 573–582.

[40] M. Jenders, G. Kasneci, and F. Naumann. Analyzing and predicting viral tweets. In Proceedings of the 22nd international conference on world wide web, 2013, pp. 657–664.

[41] Y. Zhang, Z. Xu, and Q. Yang. Predicting popularity of messages in twitter using a feature-weighted model. Режим доступа: http://www.nlpr.ia.ac.cn/2012papers/gjhy/gh154.pdf, дата обращения 20.11.2018.

[42] Q. Cao, H. Shen, K. Cen, W. Ouyang, and X. Cheng. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In Proc. of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1149–1158.

[43] C. Li, J. Ma, X. Guo, and Q. Mei. Deepcas: An end-to-end predictor of information cascades. In Proc. of the 26th International Conference on World Wide Web, 2017, pp. 577–586.

[44] B. Shulman, A. Sharma, and D. Cosley. Predictability of popularity: Gaps between prediction and understanding. in Proc. of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016), 2016, pp. 348–357.

[45] S. Petrovic, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. In Proc. of the Fifth International AAAI Conference on Weblogs and Social Media, 2011, pp. 586–589.

[46] Q. Zhang, Y. Gong, J. Wu, H. Huang, and X. Huang. Retweet prediction with attention-based deep neural network. In Proc. of the 25th ACM International on Conference on Information and Knowledge Management, 2016, pp. 75–84.

[47] B. Jiang, Z. Lu, N. Li, J. Wu, and Z. Jiang. Retweet prediction using social-aware probabilistic matrix factorization. Lecture Notes in Computer Science, vol. 10860, 2018, pp. 316–327.

[48] Y. Yang, J. Tang, C. W.-k. Leung, Y. Sun, Q. Chen, J. Li, and Q. Yang. Rain: Social role-aware information diffusion. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 367–373.

[49] S. Choobdar, P. Ribeiro, S. Parthasarathy, and F. Silva. Dynamic inference of social roles in information cascades. Data mining and knowledge discovery, vol. 29, no. 5,

2015, pp. 1152–1177.
[50] L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. International journal of emergency management, vol. 6, no. 3-4, 2009, pp. 248–260.
[51] P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Maziad. Opening closed regimes: what was the role of social media during the arab spring? Режим доступа: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2595096, дата обращения 20.11.2018.

Methods for Information Spread Analysis

[1,2] *Avetisyan A. A.<a.a.avetisyan@ispras.ru>*
[1,3] *Drobyshevskiy M. D.<drobyshevsky@ispras.ru>*
[1,2,4] *Turdakov D.Yu.<turdakov@ispras.ru>*

[1]*Ivannikov Institute for System Programming of the RAS,*
25, Alexander Solzhenitsyn Str., Moscow, 109004, Russia
[2]*Lomonosov Moscow State University,*
GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation
[3]*Moscow Institute of Physics and Technology,*
9 Institutskiy per., Dolgoprudny, Moscow Region, 141700, Russian Federation
[4] *National Research University Higher School of Economics (HSE)*
11 Myasnitskaya Ulitsa, Moscow, 101000, Russia

Abstract. Spread of information is a fundamental process taking place on the Internet. Every day we can observe the publication of various information and its further dissemination through news articles and messages from ordinary users. Although the process itself can be observed explicitly, it is difficult to determine individual propagation paths. The increase of global information in all spheres of human life radically changes the speed and ways of disseminating information. In this review, we study models of information flows on the Internet and divide them into two groups: explanatory models, which suggest the existence of the underlying network over which information propagates, and predictive models, studying spread of individual information flows. Despite all the complexity, the study of the important properties of information spread is necessary for understanding the general processes occurring in the modern information society.

Keywords: information diffusion, information cascades, networks of diffusion

DOI: 10.15514/ISPRAS-2018-30(6)-11

For citation: Avetisyan A.A., Drobyshevskiy M. D., Turdakov D.Yu. Methods for Information Spread Analysis. Trudy ISP RAN/Proc. ISP RAS, vol. 30, issue 6, 2018, pp. 199-220 (in Russian). DOI: 10.15514/ISPRAS-2018-30(6)-11

References

[1] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 1019-1028.
[2] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, and Y.-C. Zhang. Dynamics of

information diffusion and its applications on complex networks. *Physics Reports*, vol. 651, 2016, pp. 1-34.

[3] Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In Proc. of the International workshop on privacy enhancing technologies. Springer, 2006, pp. 36-58.

[4] S. Fournier and J. Avery. The uninvited brand. Business horizons, vol. 54, no. 3, pp. 193-207, 2011.

[5] G.E. Kreindler and H.P. Young. Rapid innovation diffusion in social networks. In Proceedings of the National Academy of Sciences, vol. 111, supplement 3, 2014, pp. 10 881-10 888.

[6] T. Holz, M. Steiner, F. Dahl et al. Measurements and mitigation of peer-to-peer-based botnets: A case study on storm worm. In Proc. of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, 2008.

[7] B. Doerr, M. Fouz, and T. Friedrich. Why rumors spread so quickly in social networks. Communications of the ACM, vol. 55, no. 6, 2012, pp. 70-75.

[8] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 497-506.

[9] Y. Sano, K. Yamada, H. Watanabe, H. Takayasu, and M. Takayasu. Empirical analysis of collective human behavior for extraordinary events in the Blogosphere. Physical Review E, vol. 87, no. 1, 2013.

[10] J. Lehmann, B. Goncalves, J.J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In Proc. of the 21st international conference on World Wide Web, 2012, pp. 251–260.

[11] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In Proc. of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 6–14.

[12] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In Proc. of the 20th international conference on World wide web, 2011, pp. 695–704.

[13] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H.E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. In Proceedings of the National Academy of Sciences, vol. 113, no. 3, 2016, pp. 554–559.

[14] S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. M. Tripathy, and S. Triukose. Spatio-temporal and events based analysis of topic popularity in twitter. In Proc. of the 22nd ACM international conference on Information & Knowledge Management, 2013, pp. 219–228.

[15] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In Proc. of the 21st international conference on World Wide Web, 2012, pp. 519–528.

[16] S. Goel, D.J. Watts, and D.G. Goldstein. The structure of online diffusion networks. In Proc. of the 13th ACM conference on electronic commerce, 2012, pp. 623–638.

[17] S. Goel, A. Anderson, J. Hofman, and D. J. Watts. The structural virality of online diffusion. Management Science, vol. 62, no. 1, 2015, pp. 180–196.

[18] S. Krishnan, P. Butler, R. Tandon, J. Leskovec, and N. Ramakrishnan. Seeing the forest for the trees: new approaches to forecasting cascades. In Proc. of the 8th ACM Conference on Web Science, 2016, pp. 249–258.

[19] S.A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 33–41.
[20] E. Bakshy, S. Messing, and L.A. Adamic. Exposure to ideologically diverse news and opinion on facebook. Science, vol. 348, no. 6239, 2015, pp. 1130–1132.
[21] A. Vazquez, B. Racz, A. Lukacs, and A.-L. Barabasi. Impact of non-poissonian activity patterns on spreading processes. Physical review letters, vol. 98, no. 15, 2007.
[22] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In Proc. of the 2010 IEEE 10th International Conference on Data Mining (ICDM), 2010, pp. 599–608.
[23] M.J. Salganik, P.S. Dodds, and D.J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. Science, vol. 311, no. 5762, 2006, pp. 854–856.
[24] J. Cheng, L. Adamic, P.A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In Proc. of the 23rd international conference on World wide web, 2014, pp. 925–936.
[25] MemeTracker data. Режим доступа: http://www.memetracker.org/data.html, дата обращения 20.11.2018.
[26] D. Centola, V. M. Eguıluz, and M.W. Macy. Cascade dynamics of complex propagation. Physica A: Statistical Mechanics and its Applications, vol. 374, no. 1, 2007, pp. 449–456.
[27] A. Nematzadeh, E. Ferrara, A. Flammini, and Y.-Y. Ahn. Optimal network modularity for information diffusion. Physical review letters, vol. 113, 2014, no. 8.
[28] P.S. Dodds and D.J. Watts. A generalized model of social and biological contagion. Journal of theoretical biology, vol. 232, no. 4, 2005, pp. 587–604.
[29] C.D. Brummitt, K.-M. Lee, and K.-I. Goh. Multiplexity-facilitated cascades in networks. Physical Review E, vol. 85, no. 4, 2012.
[30] F. Karimi and P. Holme. Threshold model of cascades in empirical temporal networks. Physica A: Statistical Mechanics and its Applications, vol. 392, no. 16, 2013, pp. 3476–3483.
[31] L. Dickens, I. Molloy, J. Lobo, P.-C. Cheng, and A. Russo. Learning stochastic models of information flow. In Proc. of the 2012 IEEE 28th international conference on data engineering, 2012, pp. 570–581.
[32] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 137–146.
[33] M. Kimura, K. Saito, and R. Nakano. Extracting influential nodes for information diffusion on a social network. In Proc. of the Twenty-Second AAAI conference on Artificial intelligence, vol. 2, 2007, pp. 1371–1376.
[34] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. Computer networks, vol. 33, no. 1-6, 2000, pp. 309–320.
[35] M. Kimura, K. Saito, and H. Motoda. Minimizing the spread of contamination by blocking links in a network. in Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, 2008, pp. 1175–1180.
[36] K. Lerman. Information is not a virus, and other consequences of human cognitive limits. Future Internet, vol. 8, no. 2, 2016.
[37] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal

dynamics of diffusion networks. arXiv preprint arXiv:1105.0697, 2011.

[38] M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf. Structure and dynamics of information pathways in online media. In Proc. of the sixth ACM international conference on Web search and data mining, 2013, pp. 23–32.

[39] S. Bourigault, S. Lamprier, and P. Gallinari. Representation learning for information diffusion through social networks: an embedded cascade model. In Proc. of the Ninth ACM International Conference on Web Search and Data Mining, 2016, pp. 573–582.

[40] M. Jenders, G. Kasneci, and F. Naumann. Analyzing and predicting viral tweets. In Proceedings of the 22nd international conference on world wide web, 2013, pp. 657–664.

[41] Y. Zhang, Z. Xu, and Q. Yang. Predicting popularity of messages in twitter using a feature-weighted model. Available at: http://www.nlpr.ia.ac.cn/2012papers/gjhy/gh154.pdf, accessed 20.11.2018.

[42] Q. Cao, H. Shen, K. Cen, W. Ouyang, and X. Cheng. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In Proc. of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1149–1158.

[43] C. Li, J. Ma, X. Guo, and Q. Mei. Deepcas: An end-to-end predictor of information cascades. In Proc. of the 26th International Conference on World Wide Web, 2017, pp. 577–586.

[44] B. Shulman, A. Sharma, and D. Cosley. Predictability of popularity: Gaps between prediction and understanding. in Proc. of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016), 2016, pp. 348–357.

[45] S. Petrovic, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. In Proc. of the Fifth International AAAI Conference on Weblogs and Social Media, 2011, pp. 586–589.

[46] Q. Zhang, Y. Gong, J. Wu, H. Huang, and X. Huang. Retweet prediction with attention-based deep neural network. In Proc. of the 25th ACM International on Conference on Information and Knowledge Management, 2016, pp. 75–84.

[47] B. Jiang, Z. Lu, N. Li, J. Wu, and Z. Jiang. Retweet prediction using social-aware probabilistic matrix factorization. Lecture Notes in Computer Science, vol. 10860, 2018, pp. 316–327.

[48] Y. Yang, J. Tang, C. W.-k. Leung, Y. Sun, Q. Chen, J. Li, and Q. Yang. Rain: Social role-aware information diffusion. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 367–373.

[49] S. Choobdar, P. Ribeiro, S. Parthasarathy, and F. Silva. Dynamic inference of social roles in information cascades. Data mining and knowledge discovery, vol. 29, no. 5, 2015, pp. 1152–1177.

[50] L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. International journal of emergency management, vol. 6, no. 3-4, 2009, pp. 248–260.

[51] P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Maziad. Opening closed regimes: what was the role of social media during the arab spring? Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2595096, accessed 20.11.2018.