ФГБОУ ВПО «Воронежский государственный
технический университет»

Е.А. Шварцкопф О.А. Остапенко

# СОЦИАЛЬНЫЕ СЕТИ: РИСКИ И ОБЕСПЕЧЕНИЕ БЕЗОПАСНОСТИ

Утверждено Редакционно-издательским советом
университета в качестве учебного пособия

Воронеж 2015

УДК 004.056

Шварцкопф Е. А. Социальные сети: риски и обеспечение безопасности: учеб. пособие [Электронный ресурс]. – Электрон. текстовые, граф. данные (1,27 Мб) / Е. А. Шварцкопф, О. А. Остапенко. – Воронеж: ФГБОУ ВПО «Воронежский государственный технический университет», 2015. – 1 электрон. опт. диск (CDROM). – Систем. требования: ПК 500 и выше; 256 Мб ОЗУ; Windows XP; Adobe Reader; 1024x768; CD-ROM; мышь. – Загл. с экрана.

В пособии рассматриваются проблемы социальных сетей с точки зрения повышения защищённости пользователей социальных сетей путём анализа моделей распространения вредоносного программного обеспечения, а также с помощью построения риск-модели информационно-психологического воздействия на пользователей социальных информационных сетей.

Издание соответствует требованиям Федерального государственного образовательного стандарта высшего профессионального образования по специальности 090303 «Информационная безопасность автоматизированных систем», дисциплине «Социальные сети: риски и обеспечение безопасности».

Табл. 2. Ил. 27. Библиогр.: 72 назв.

Рецензенты: ОАО «Концерн «Созвездие»
(канд. техн. наук, ведущий науч. сотрудник О.В. Поздышева);
д-р техн. наук, проф. А.Г. Остапенко

# ВВЕДЕНИЕ

В XX веке в силу информационной революции технические подсистемы носят все более кибернетический характер, их основой во многом является наполнитель - информация. Информация и информационное пространство стали системообразующим фактором для отдельных социумов и общества в целом.

В XXI веке, начало которого мы переживаем, человеко-машинная модель развития становится нормой для личности, общества и государства. Люди понуждают киберсистемы на производство и обработку информации, которую использует социум. При этом процесс создания я потребления информации носит обязательный характер, и теперь он уже жизненно необходим для общества, государства и личности.

Следует заметить, что инфраструктура таких информационных отношений носит все более разветвлённый характер. Причем, интеллектуализация процессов генерации новой информации та всех уровнях привела к широкому распространению распределенных систем, где имеет место пространственная и временная рассредоточенность ресурсов и функций между компонентами системы, коммутируемых ею через соответствующие информационные потоки. В данном случае имеется в виду не только и не столько физическая рассредоточенность компонентов, сколько их централизованная управляемость и относительная автономность в информационном пространстве, в котором функционируют современные социальные информационные системы (СИС).

Глобализация информационного пространства естественно усиливает рассматриваемую тенденцию. Современные информационно-кибернетические сети, прежде всего Интернет, обусловливают мульти размерность и высокую коммуникабельность элементов множества субъектов информационных отношений в СИС: генераторов, транспортеров и пользователей информации.

Современные военные кампании также сопровождаются информационными операциями и атаками (ИОА) в открытых телекоммуникационных системах и средствах массовой информации (СМИ). При этом ИОА определяют как операции, связанные с информацией, которые заправлены на достижение превосходства в информационной сфере, т.е. в ИКП и/или ИПП. В связи с этим некоторые горячие головы рассматривают совокупность информационно- кибернетических атак (ИКА) и информационно-психологических операций (ИПО) как третью мировую войну - информационную войну (ИВ) между супер социотехническими системами посредством ИОА, реализуемых средствами ИКТ.

С середины XX века, пожалуй, самым ожесточенным полем сражения ИВ стали кибернетические системы. Компьютерный разбой и электронный шпионаж стали расхожими средствами конкурентной борьбы. Проблема радиоэлектронной борьбы и технической защиты информации была возведена в ранг государственной. Для автоматизированных систем в защищенном исполнении получили развитие криптографические, правовые, организационные, программные и аппаратные средства противодействия ИКА. Реальной угрозой обществу стала преступность в сфере высоких технологий, посягательства на служебную и коммерческую тайну.

На старте XXI века сражения ИВ развернулись за человеческие умы. Правительственные структуры, СМИ и неформальные движения включились в ИПО с целью управления общественным мнением.

# 1. МЕТОДЫ АНАЛИЗА КОМПЬЮТЕРНЫХ СОЦИАЛЬНЫХ СЕТЕЙ

## 1.1. Основные направления исследования компьютерных социальных сетей

В настоящее время в анализе социальных сетей выделяют [1] четыре основных направления исследований: структурное, ресурсное, нормативное и динамическое. В *структурном подходе* все участники сети рассматриваются как вершины графа, которые влияют на конфигурацию ребер и других участников сети. Основное внимание уделяется геометрической форме сети и интенсивности взаимодействий (весу ребер), поэтому исследуются такие характеристики, как взаимное расположение вершин, центральность, транзитивность взаимодействий. Для интерпретации результатов в данном направлении используются структурные теории и теории сетевого обмена.

*Ресурсный подход* рассматривает возможности участников по привлечению индивидуальных и сетевых ресурсов для достижения определенных целей и дифференцирует участников, находящихся в идентичных структурных позициях социальной сети, по их ресурсам. В качестве индивидуальных ресурсов могут выступать знания, престиж, богатство, раса, пол. Под сетевыми ресурсами понимаются влияние, статус, информация, капитал.

*Нормативное направление* изучает уровень доверия между участниками, а также нормы, правила и санкции, которые влияют на поведение участников в социальной сети и процессы их взаимодействий. В этом случае анализируются социальные роли, которые связаны с данным ребром сети, например, отношения руководителя и подчиненного, дружеские или родственные связи. Комбинация индивидуальных и сетевых ресурсов участника с нормами и правилами, действующими в данной социальной сети, образует его «сетевой капитал». В упрощенном виде «сетевой капитал» можно рассматривать как сумму некоторых преимуществ,

которые участник может получить в произвольный момент времени для достижения некоторой цели.

*Динамический подход* – направление в изучении социальных сетей, в котором объектами исследований являются изменения в сетевой структуре с течением времени: по каким причинам исчезают и появляются ребра сети, как сеть изменяет свою структуру при внешних воздействиях, существуют ли какие-либо стационарные конфигурации социальной сети и др. Рассмотрим немного подробнее перечисленные направления анализа социальных сетей в терминах решаемых задач [2].

*Структурный анализ и анализ поведения связей* в социальных сетях необходим для того, чтобы определить наиболее важные вершины, связи, сообщества и развивающиеся регионы сети. Такой анализ позволяет осуществлять обзор глобального эволюционного поведения сети. При структурном анализе и анализе поведения связей используются методы статистического анализа, методы определения сообществ, алгоритмы классификации.

*Статистический анализ социальных сетей.* В работе [3] приведен развернутый анализ структурных свойств сетей большого размера. Исследуются «типичные» социальные сети и изучается вопрос, как будет выглядеть сеть, если ее увеличить. Изучается взаимное поведение вершин сети исходя из предположения, что у большинства вершин имеется мало связей, возникают ли при этом «ядра» (скопления) или степени вершин распределяются более равномерно. Изучается поведение вершин при кластеризации. Другой из вопросов– поведение типичных временных характеристик социальных сетей. Например, как меняется структура сети в процессе роста или как меняется поведение и распределение связных компонентов графа. Со временем к сети добавляются новые сущности, но, несмотря на это, некоторые свойства графа могут сохраняться.

*Определение сообществ в социальных сетях.* Этот вопрос является наиболее важным в анализе социальных сетей,

хотя довольно близок к задаче классификации. Цель – попытаться определить регионы сети, внутри которых происходит активное взаимодействие участников. Алгоритмически эту задачу можно отнести к задаче о разделении графов. Нужно разделить сеть на плотные регионы на основе поведения связей между вершинами. Компьютерные социальные сети динамические, что приводит к затруднениям с точки зрения выявления сообществ. В некоторых случаях удается интегрировать информационное содержимое сети в процесс определения сообществ. Тогда контент является вспомогательным средством для выявления групп участников с похожими интересами.

*Анализ содержания социальных сетей.* Можно выделить четыре вида анализа контента сети:

- анализ общей информации с произвольными типами данных;
- анализ текста;
- анализ мультимедиа;
- сенсорный и потоковый анализ.

*Анализ медиаданных.* Для обнаружения полезных бизнес-приложений можно анализировать социальные медиасети. Техники анализа данных предоставляют исследователям и специалистам инструменты для анализа больших, комбинированных, постоянно меняющихся медиаданных.

*Анализ текстовой информации в социальных сетях.* В вершинах социальной сети содержится много текстовой информации в различных формах, например, ссылки на посты (сообщения), блоги или статьи с новостями. Иногда пользователи могут отмечать друг друга, что тоже является формой текстовой информации в виде ссылок. Использование контента сети может сильно улучшить качество выводов при анализе социальных сетей, например, в задачах кластеризации и классификации.

*Интеграция данных, поступающих с датчиков, и социальных сетей.* Многие современные сотовые телефоны

поддерживают возможность взаимодействия пользователей друг с другом динамически в режиме реального времени в зависимости от их местоположения и статуса.

Подобные приложения также приводят к образованию потоков массивов в режиме реального времени. Их применяют для того, чтобы получить информацию о человеке или совокупности свойств объектов, которые отслеживаются. Поскольку информация о местоположении пользователя является личной, естественно возникает ряд проблем с точки зрения обработки, исследуются методы интеграции данных, поступающих с датчиков, и данных в социальных сетях.

*Анализ мультимедийной информации сети.* Существует много сайтов по обмену средствами массовой информации, например, Flickr и YouTube, которые обеспечивают возможность совместного использования этой информации. Такие средства массовой информации общего пользования часто сочетаются с взаимодействием пользователей – размещением тегов и комментариев в различных изображениях. Поэтому подобные сети могут служить источником для широкого спектра приложений в процессе извлечения и анализа данных.

*Расстановка тегов.* Большинство взаимодействий между пользователями происходит в форме тегирования (расстановки тегов, отметок), в которых пользователи прикрепляют описания различных объектов в социальной сети, такие как картинки, текст, видео или другая мультимедийная информация. В рамках данного подхода изучают свойства потоков тегов, моделей тегирования, семантику тегов, рекомендации по использованию тегов, визуализацию тегов, приложения для расстановки тегов, интеграцию различных систем тегирования и проблем, связанных с использованием тегов. Интересным вопросом является, например, почему люди расставляют теги, что влияет на выбор при тегировании, как промоделировать процесс подобной разметки, разновидности тегов, как создаются теги и как выбрать правильные теги для рекомендации.

*Случайные блуждания и их применение в социальных сетях.* Классификация – один из наиболее известных методов в веб-поиске. Например, можно упомянуть алгоритм ссылочного ранжирования (PageRank) для приписывания веса веб-документам. Его основной принцип может применяться для поиска и классификации сущностей и участников в социальной сети. Этот алгоритм использует подход случайного блуждания для того, чтобы оценить вероятность посещения той или иной вершины. Естественно, что вершины, которые лучше расположены со структурной точки зрения, имеют более высокий вес, а значит, являются более важными. Методы случайного блуждания могут быть также полезны для объединения участников в группы относительно наиболее влиятельных участников.

*Классификация вершин в социальных сетях.* Некоторые вершины удобно снабжать пометками, чтобы их отличительные особенности и структурную информацию можно было распространить на всю сеть. Например, в маркетинговых исследованиях определенные вершины могут обозначать заинтересованность участников сети в конкретном продукте, и было бы желательно применить характерные особенности этих вершин для изучения других участников на предмет заинтересованности этим продуктом. Для этих целей, кроме того, можно использовать информацию о контенте и структуре социальной сети. Другой пример, когда об одной из двух связных вершин получены некоторые сведения, для второй эти сведения с большой долей вероятности тоже будут верны. Вот почему структуру связей можно применять для распространения меток среди вершин. Содержимое сети и структурные особенности в дальнейшем могут пригодиться для подтверждения качества полученной классификации.

*Анализ социального влияния.* Так как в основе социальных сетей лежит взаимодействие применение «вирусного маркетинга» для распространения сообщения между взаимосвязанными участниками через всю сеть. Вопросы этого направления:

- как моделировать влияние на основе информации об участниках;
- как моделировать распространение влияния;
- кто из участников является наиболее влиятельным в процессе распространения?

*Конфиденциальность в социальных сетях.* В социальных сетях содержится большое количество личной информации об участниках, например, интересы, информация о дружбе, демографическая информация и др. Это может привести к несанкционированному распространению личной информации в сетях. В решении такого типа задач полезно применять модели на основе механизмов конфиденциальности.

*Обнаружение экспертов в сетях.* Социальная сеть может являться инструментом для выявления экспертов в конкретной области. Часто в реальности эксперты образуют сеть, которая соответствует социальной сети или организационной структуре компании. Многие сложные задачи требуют коллективного решения нескольких экспертов. В подобных случаях получается, что более эффективно можно достигнуть общую цель, когда специалисты сотрудничают друг с другом. Важной задачей данного направления является обнаружение групп специалистов в определенных узких областях.

*Эволюция в динамических социальных сетях.* С течением времени в социальных сетях появляются новые участники, некоторые участники прекращают взаимодействие, возникают новые связи, некоторые связи устаревают, так как участники перестают взаимодействовать. Это приводит к изменениям в структуре социальных сетей в целом и в отдельных сообществах. При этом возникает два важных вопроса: 1) согласно каким законам происходят долгосрочные изменения между крупными сообществами в социальных сетях; 2) как развиваются сами сообщества во времени. Какие изменения могут происходить, как можно отследить и представить их?

*Прогноз формирования связей в социальных сетях.* Для извлечения интересующей информации из социальной сети полезны исследования, направленные на определение и предсказание возможных связей между вершинами в будущем.

В большинстве приложений для анализа социальных сетей связи считаются динамическими и могут сильно изменяться с течением времени. Например, отношение «дружбы» не меняется. В процесс прогнозирования связей может быть вовлечена как структура сети, так и информация об особенностях различных вершин. Для решения таких задач предлагается строить разнообразные структурные и реляционные модели.

*Визуализация социальных сетей.* Социальные сети становятся крупнее и имеют все более сложную структуру. Визуализация помогает естественным образом свести воедино информацию о сетях и сделать ее более доступной для понимания. Визуализация в сочетании с взаимодействием помогают аналитикам в описании социальных сетей. Целью этого направления является также поиск ответа на вопрос, как различные модели могут быть использованы для изучения различных аспектов сетей, таких как структура и семантика. Важным является создание алгоритмов, сочетающих методы анализа и визуализации, чтобы улучшить понимание структуры и динамики сети. Как видно, при анализе социальных сетей решается довольно большой круг задач и применяются методы из различных областей знаний.

## 1.2. Некоторые наиболее известные социальные сети

К крупнейшим социальным сетям по числу пользователей относятся Facebook, LinkedIn, ВКонтакте, Twitter, Одноклассники.ru, YouTube и др.

*Facebook* (http://www.facebook.com). Сеть основана в 2004 г. Марком Цукербергом. По данным за апрель 2012 г. аудитория Facebook составляет 901 млн пользователей. Каждый день в сети пользователи оставляют 3,2 млрд «лайков»

и комментариев и публикуют 300 млн фотографий. Facebook позволяет создать профиль с фотографией и информацией о себе, приглашать друзей, обмениваться с ними сообщениями, изменять свой статус, оставлять сообщения на своей и чужой «стенах», загружать фотографии и видеозаписи, создавать группппы (сообщества по интересам). Существует возможность создавать приложения для Facebook (игры, средства обмена музыкой, фотографиями и т.д.), что повышает посещаемость сайта.

*YouTube* (http://www.youtube.com). Сервис, предоставляющий услуги видеохостинга, основан в 2005 г. Пользователи могут добавлять, просматривать и комментировать те или иные видеозаписи, добавлять аннотации и титры к видео, а также выставлять рейтинг просмотренным видео, если такую возможность им предоставил автор. Благодаря простоте и удобству использования YouTube стал популярнейшим видеохостингом и третьим сайтом в мире по количеству посетителей на июнь 2012 г. Ежеминутно на YouTube загружают 60 часов видео. В январе 2012 г. ежедневное количество просмотров видео на сайте достигло 4 млрд.

*LinkedIn* (http://www.linkedin.com). Социальная сеть была основана Ридом Хоффманом в декабре 2002 г., запущена в мае 2003 г. В основном сеть используется для поиска и установления деловых контактов. По данным на февраль 2012 г. 4, в LinkedIn зарегистрировано свыше 160 млн пользователей. Чуть меньше половины пользователей LinkedIn являются жителями США.

*ВКонтакте* (http://vk.com). Сеть основана в 2006 г. Павлом Дуровым. По данным на март 2012 г. аудитория ВКонтакте составила около 150 млн человек, около 70 % из них проживают в России. Подобно Facebook, пользователи ВКонтакте могут обмениваться сообщениями приватно (через личные сообщения) и публично (с помощью записей на «стене», а также через механизм групп и встреч), отслеживать через ленту новостей активность друзей и сообществ. В сети есть возможность обмена и загрузки файлов довольно

большого объема, так как используется технология распределенного распространения файлов BitTorrent, что делает ВКонтакте одним из крупнейших медиаархивов Рунета. Facebook, Одноклассники.ru и другие социальные сети использует протокол обмена сообщениями XMPP (Extensible Messaging and Presence Protocol), ранее известный как Jabber.

*Twitter* (https://twitter.com). Создана в 2006 г. Джеком Дорси. По состоянию на начало 2012 г. сервис насчитывает более 140 млн пользователей. Ежедневно пользователи отправляют около 340 млн сообщений. Система позволяет отправлять короткие текстовые сообщения (до 140 символов), используя веб-интерфейс, SMS, средства мгновенного обмена сообщениями или сторонние программы-клиенты. Отличительной особенностью Twitter является публичная доступность размещенных сообщений, что позволяет называть его микроблогом. С 2011 г. Twitter перешел с MySQL на Lucene и с Ruby on Rails на Java и Scala для повышения производительности и масштабируемости.

Одноклассники.ru (http://www.odnoklassniki.ru). Проект запущен в 2006 г., его автором является российский веб-разработчик Альберт Попков. По состоянию на июнь 2011 г. зарегистрировано более 70 млн пользователей. Особенность этой сети в том, что каждый пользователь видит имена всех, кто заходил посмотреть на его анкету, все публичные действия пользователей (сообщения в форумах, добавление друзей, загрузка фотографий) отображаются в доступной другим пользователям «ленте активности». Является русскоязычным аналогом американской сети Classmates.com.

*Flickr* (www.flickr.com). Создана в 2004 г. По данным на июнь 2011 г. в сети зарегистрировано 51 млн пользователей. Сервис предназначен для хранения и дальнейшего использования пользователем цифровых фотографий и видеороликов. Является одним из первых Web 2.0 сервисов. Есть возможность к каждой фотографии добавить название, краткое описание и ключевые слова (тег) для дальнейшего поиска.

## 1.3. Параметры сложных сетей

В теории сложных сетей выделяют три основных направления: исследование статистических свойств, которые характеризуют поведение сетей; создание модели сетей; предсказание поведения сетей при изменении структурных свойств. В прикладных исследованиях обычно применяют такие типичные для сетевого анализа характеристики, как размер сети, сетевая плотность, степень центральности и т. п. При анализе сложных сетей как и в теории графов исследуются параметры отдельных узлов; параметры сети в целом; сетевые подструктуры.

### 1.3.1. Параметры узлов сети

Для отдельных узлов выделяют следующие параметры:

– входная степень узла – количество ребер графа, которые входят в узел;

– выходная степень узла – количество ребер графа, которые выходят из узла;

– расстояние от данного узла до каждого из других;

– среднее расстояние от данного узла до других;

– эксцентричность (eccentricity) – наибольшее из геодезических расстояний (минимальных расстояние между узлами) от данного узла к другим;

– посредничество (betwetnness), показывающее, сколько кратчайших путей проходит через данный узел;

– центральность – общее количество связей данного узла по отношению к другим.

### 1.3.2. Общие параметры сети

Для расчета индексов сети в целом используют такие параметры, как: число узлов, число ребер, геодезическое расстояние между узлами, среднее расстояние от одного узла к другим, плотность – отношение количества ребер в сети к возможному максимальному количеству ребер при данном количестве узлов, количество симметричных, транзитивных и циклических триад, диаметр сети – наибольшее геодезическое расстояние в сети и т.д..

Существует несколько актуальных задач исследования сложных сетей, среди которых можно выделить следующие основные:

– определение клик в сети. Клики – это подгруппы или кластеры, в которых узлы связаны между собой сильнее, чем с членами других клик;

– выделение компонент (частей сети), которые связаны внутри и не связаны между собой;

– нахождение блоков и перемычек. Узел называется перемычкой, если при его изъятии сеть распадается на несвязанные части;

– выделение группировок – групп эквивалентных узлов (которые имеют максимально похожие профили связей).

### 1.3.3. Распределение степеней узлов

Важной характеристикой сети является функция распределения степеней узлов которая определяется как вероятность того, что узел имеет степень . Сети, характеризующиеся разными , демонстрируют весьма разное поведение. в некоторых случаях может быть распределениями Пуассона ( ——— где математическое ожидание), экспоненциальным ( ) или степенным ( — ).

Сети со степенным распределением степеней узлов называются безмасштабными (scale-free). Именно безмасштабные распределения часто наблюдаются в реально существующих сложных сетях. При степенном распределении возможно существование узлов с очень высокой степенью, что практически не наблюдается в сетях с пуассоновым распределением.

### 1.3.4. Путь между узлами

Расстояние между узлами определяется как количество шагов, которые необходимо сделать, чтобы по существующим ребрам добраться от одного узла до другого. Естественно, узлы могут быть соединены прямо или опосредованно. Путем между узлами назовем кратчайшее расстояние между ними. Для всей сети можно ввести понятие среднего пути, как среднее по всем парам узлов кратчайшего расстояния между ними:

$$\overline{\qquad}$$

где – количество узлов, кратчайшее расстояние между узлами и .

Венгерскими математиками П. Эрдёшем (P. Erdős) и А. Реньи (A. Rényi) было показано, что среднее расстояние между двумя вершинами в случайном графе растет как логарифм от числа вершин [36,37].

С именем П. Эрдёша связаны не только исследования сложных сетей, но и популярное число Эрдёша, которое используется как один из критериев определения уровня математиков в соответствующем социуме, базирующийся на так называемой сети соавторства. Известно, что Эрдёш написал около полутора тысяч статей, а также, что количество его соавторов превышало 500. Столь большое число соавторов и породило такое понятие, как число Эрдёша, которое

определяется следующим образом: у самого Эрдёша это число равно нулю; у соавторов Эрдёша это число равно единице; соавторы людей с числом Эрдёша, равным единице, имеют число Эрдёша два; и так далее.

Таким образом, число Эрдёша это длина пути от некоторого автора до самого Эрдёша по совместным работам. Известен факт, что 90% математиков обладают числом Эрдёша не выше 8, что соответствует теории «малых миров», речь о которой пойдет ниже.

Некоторые сети могут оказаться несвязными, т.е. найдутся узлы, расстояние между которыми окажется бесконечным. Соответственно, средний путь может оказаться также равным бесконечности. Для учета таких случаев вводится понятие среднего инверсного пути между узлами, рассчитываемое по формуле:

Сети также характеризуются таким параметром как диаметр или максимальный кратчайший путь, равный максимальному значению из всех

### 1.3.5. Коэффициент кластерности

Д. Уаттс (D. Watts) и С. Строгатц (S. Strogatz) в 1998 году определили такой параметр сетей, как коэффициент кластерности [38], который соответствует уровню связности узлов в сети. Этот коэффициент характеризует тенденцию к образованию групп взаимосвязанных узлов, так называемых клик (clique). Кроме того, для конкретного узла коэффициент кластеризации показывает, сколько ближайших соседей данного узла являются также ближайшими соседями друг для друга.

Коэффициент кластерности для отдельного узла сети определяется следующим образом. Пусть из узла выходит $k$ ребер, которые соединяют его с $k$ другими узлами, ближайшими соседями. Если предположить, что все ближайшие соседи соединены непосредственно друг с другом, то количество ребер между ними составляло бы – . То есть это число, которое соответствует максимально возможному количеству ребер, которыми могли бы соединяться ближайшие соседи выбранного узла. Отношение реального количества ребер, которые соединяют ближайших соседей данного узла к максимально возможному (такому, при котором все ближайшие соседи данного узла были бы соединены непосредственно друг с другом) называется коэффициентом кластерности узла Естественно, эта величина не превышает единицы.

Коэффициент кластерности может определяться как для каждого узла, так и для всей сети. Соответственно, уровень кластерности всей сети определяется как нормированная по количеству узлов сумма соответствующих коэффициентов отдельных узлов. Рассмотренный ниже феномен «малых миров» непосредственно связан с уровнем кластерности сети.

### 1.3.6. Посредничество

Посредничество (betweenness) – это параметр, показывающий, сколько кратчайших путей проходит через узел. Эта характеристика отражает роль данного узла в установлении связей в сети. Узлы с наибольшим посредничеством играют главную роль в установлении связей между другими узлами в сети. Посредничество узла определяется по формуле:

где – общее количество кратчайших путей между узлами и ;

– количество кратчайших путей между узлами и ,проходящих через узел .

### 1.3.7. Эластичность сети

Свойство эластичности сетей относится к распределению расстояний между узлами при изъятии отдельных узлов. Эластичность сети зависит от ее связности, т.е. существовании путей между парами узлов. Если узел будет изъят из сети, типичная длина этих путей увеличится. Если этот процесс продолжать достаточно долго, сеть перестанет быть связной. Р. Альберт (Réka Albert) из университета штата Пенсильвания, США при исследовании атак на интернет-серверы изучала эффект изъятия узла сети, представляющей собой подмножество WWW из 326000 страниц [39].

Среднее расстояние между двумя узлами, как функция от количества изъятых узлов, почти не изменилось при случайном удалении узлов (высокая эластичность). Вместе с тем целенаправленное удаление узлов с наибольшим количеством связей приводит к разрушению сети. Таким образом, Интернет является высоко эластичной сетью по отношению к случайному отказу узла в сети, но высокочувствительной к намеренной атаке на узлы с высокими степенями связей с другими узлами.

### 1.3.8. Структура сообщества

О «структуре сообщества» можно говорить тогда, когда существуют группы узлов, которые имеют высокую плотность ребер между собой, при том, что плотность ребер между отдельными группами – низкая. Традиционный метод для выявления структуры сообществ – кластерный анализ. Существуют десятки приемлемых для этого методов, которые базируются на разных мерах расстояний между узлами,

взвешенных путевых индексах между узлами и т.п. В частности, для больших социальных сетей наличие структуры сообществ оказалось неотъемлемым свойством.

## 1.4. Модели анализа социальных сетей

Один из самых известных примеров анализа сетей был проведен в 1970-е гг. американским социологом Марком Грановеттером [4]. Он показал, что для многих социальных задач, таких как поиск работы, слабые связи оказываются намного эффективнее, чем сильные. Этот эффект он назвал «*силой слабых связей*».

Некоторые свойства реальных сетей не укладываются в рамки традиционных моделей. К таким свойствам относятся и так называемые «слабые» связи. Аналогом слабых социальных связей являются, например, отношения с далекими знакомыми и коллегами. В некоторых случаях эти связи оказываются более эффективными, чем связи «сильные». Так, группой исследователей из Великобритании, США и Венгрии, был получен концептуальный вывод в области мобильной связи, заключающийся в том, что «слабые» социальные связи между индивидуумами оказываются наиболее важными для существования социальной сети [40].

Для исследования были проанализированы звонки 4.6 млн. абонентов мобильной связи, что составляет около 20% населения одной европейской страны. Это был первый случай в мировой практике, когда удалось получить и проанализировать такую большую выборку данных, относящихся к межличностной коммуникации.

В социальной сети с 4.6 млн. узлов было выявлено 7 млн. социальных связей, т.е. взаимных звонков от одного абонента другому и обратно, если обратные звонки были сделаны на протяжении 18 недель. Частота и продолжительность разговоров использовались для того, чтобы определить силу каждой социальной связи.

Было выявлено, что именно слабые социальные связи (один-два обратных звонка на протяжении 18 недель) связывают воедино большую социальную сеть. Если эти связи проигнорировать, то сеть распадется на отдельные фрагменты. Если же не учитывать сильных связей, то связность сети нарушится (рис. 1.1). Оказалось, что именно слабые связи являются тем феноменом, который связывает сеть в единое целое. Надо полагать, что данный вывод справедлив и для веб-пространства, хотя исследований в этой области до сих пор не проводилось.

1 2 3

Рис. 1.1. Структура сети:
1) полная карта сети социальных коммуникаций;
2) социальная сеть, из которой изъятые слабые связи;
3) сеть, из которой изъятые сильные связи: структура сохраняет связность

Еще один широко известный пример анализа социальных сетей – эксперимент американского психолога Милгрэма [3], проведенный в 1969 г. Этот эксперимент получил название «*феномен малого мира*» (*Milgram's small world experiment*), или «*теория шести рукопожатий*» (*Six degrees of separation*). Гипотеза заключается в том, что каждый человек знаком с любым другим жителем планеты через цепочку общих знакомых, в среднем состоящую из шести человек. Пока что это утверждение не было опровергнуто.

Наоборот, в качестве доказательства правильности гипотезы выдвигается наблюдение, что диаметр большинства сетей относительно небольшой.

Несмотря на огромные размеры некоторых сложных сетей, во многих из них (и в WWW, в частности) существует сравнительно короткий путь между двумя любыми узлами – геодезическое расстояние. В 1967 г. психолог С. Милгран в результате проделанных масштабных экспериментов вычислил, что существует цепочка знакомств, в среднем длиной шесть, практически между двумя любыми гражданами США [113].

Д. Уаттс и С. Строгатц обнаружили феномен, характерный для многих реальных сетей, названный эффектом малых миров (Small Worlds) [41]. При исследовании этого феномена ими была предложена процедура построения наглядной модели сети, которой присущ этот феномен. Три состояния этой сети представлены на рис. 1.2: регулярная сеть – каждый узел которой соединен с четырьмя соседними, та же сеть, у которой некоторые «ближние» связи случайным образом заменены «далекими» (именно в этом случае возникает феномен «малых миров») и случайная сеть, в которой количество подобных замен превысило некоторый порог.

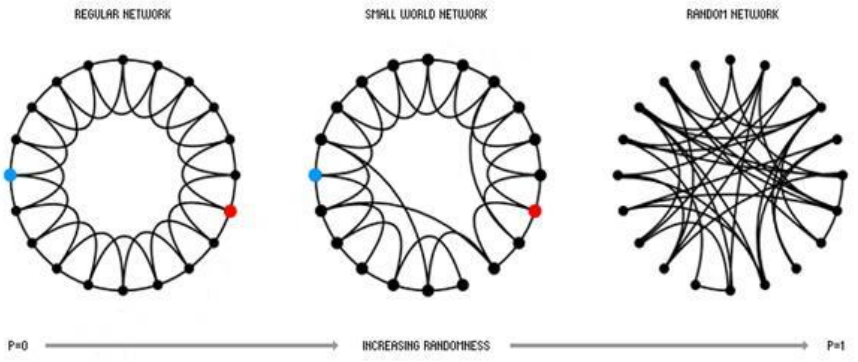



Рис. 1.2. Модель Уаттса-Строгатца

На рис. 1.3 приведены графики изменения средней длины пути и коэффициента кластеризации искусственной сети Д. Уаттса и С. Строгатца от вероятности установления «далеких связей» (в полулогарифмической шкале).

В реальности оказалось, что именно те сети, узлы которых имеют одновременно некоторое количество локальных и случайных «далеких» связей, демонстрируют одновременно эффект малого мира и высокий уровень кластеризации.

WWW является сетью, для которой также подтвержден феномен малых миров. Анализ топологии веб, проведенный Ши Жоу (S. Zhou) и Р. Дж. Мондрагоном (R.J. Mondragon) из Лондонского университета, показал, что узлы с большой степенью исходящих гиперссылок имеют больше связей между собой, чем с узлами с малой степенью, тогда как последние имеют больше связей с узлами с большой степенью, чем между собой. Этот феномен был назван «клубом богатых» (rich-club phenomenon). Исследование показало, что 27% всех соединений имеют место между всего 5% наибольших узлов, 60% приходится на соединение других 95% узлов с 5% наибольших и только 13% – это соединение между узлами, которые не входят в лидирующие 5%.

Эти исследования дают основания полагать, что зависимость WWW от больших узлов значительно существеннее, чем предполагалось ранее, т.е. она еще более чувствительна к злонамеренным атакам. С концепцией «малых миров» связан также практический подход, называемый «сетевой мобилизацией», которая реализуется над структурой «малых миров». В частности, скорость распространения информации благодаря эффекту «малых миров» в реальных сетях возрастает на порядки по сравнению со случайными сетями, ведь большинство пар узлов реальных сетей соединены короткими путями.

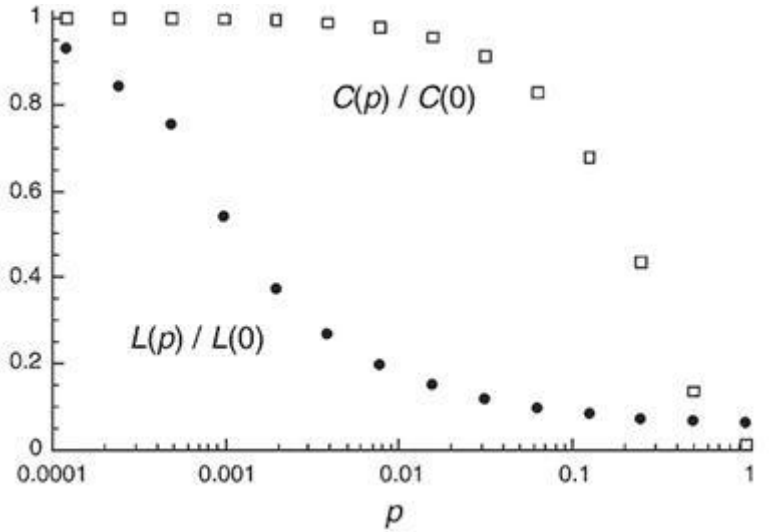



Рис. 1.3. Динамика изменения длины пути и коэффициента кластерности в модели Уаттса-Строгатца в полулогарифмической шкале (ось 0X – вероятность замены ближних связей далекими)

Кроме того, сегодня довольно успешно изучаются масштабируемые, статические, иерархические «малые миры» и другие сети, исследуются их фундаментальные свойства, такие, как стойкость к деформациям и перколяция. Недавно было показано, что наибольшую информационную проводимость имеет особый класс сетей, называемых «запутанными» (entangled networks). Они характеризуются максимальной однородностью, минимальным расстоянием между любыми двумя узлами и очень узким спектром основных статистических параметров. Считается, что запутанные сети могут найти широкое применение в области информационных технологий, в частности, в новых поколениях веб, позволяя существенным образом снизить объемы сетевого трафика.

Среди методов анализа социальных сетей основными [1] являются: методы теории графов, в частности изучение ориентированных графов и представляющих их матриц, применяемых для изучения структурных взаимосвязей

участника сети; методы нахождения локальных свойств участников, например, центральности, влиятельности, положения, принадлежности к некоторым подгруппам; методы определения эквивалентности участников, включая их структурную эквивалентность; блоковые модели и ролевые алгебры; анализ диад и триад; вероятностные модели, включая модели марковских процессов.

*Графовые модели.* Любую социальную сеть можно математически представить в виде графа

где $V$ – множество вершин графа;

$E$ – множество ребер графа;

– количество вершин в графе.

В графе социальной сети вершинами являются участники, а ребра означают наличие отношений между ними. Отношения могут быть как направленными, так и ненаправленными.

Как правило, выделяют два типа отношений: «дружба» (люди знакомы друг с другом) и «интересы» (есть общие интересы, люди входят в одну группу по интересам). Эти отношения используются, например, в FOAF (Friend of a fried) – онтологии описания людей, их активности и отношений к другим людям и объектам. В FOAF описание социальных связей между людьми основывается на транзитивности доверия. Описание алгоритма вычисления уровня доверия (TrustRank) приведено ниже.

Можно выделить три типа графовых моделей [1].

*Стохастические блоковые модели* задаются матрицей $A$ размера $N\times N$, где $N$ – число групп (блоков) участников. Элемент $\in [0,1]$ показывает плотность связей между участниками сети, принадлежащими к группе , и участниками, принадлежащими к группе При этом граф не

содержит дополнительных ребер и вершин, соответствующих связям участников внутри одной группы.

*Вероятностные графовые модели* задаются матрицей $A$ размера $N\times N$, где $N$ – число участников сети. Элемент $\in [0,1]$ показывает вероятность взаимодействия участника и участника в течение определенного периода времени.

*Обычные графовые модели* задаются матрицей связности $A$ размера $N\times N$. Для анализа графовых моделей социальных сетей иногда удобно использовать коэффициент плотности, определенный как отношение числа ребер в анализируемом графе к числу ребер в полном графе с тем же числом вершин (полный граф – это граф, в котором все вершины соединены между собой). Кроме этого, сеть могут характеризовать такие величины, как число путей заданной длины (путь – последовательность вершин, связанных между собой), минимальное число ребер, удаление которых разбивает граф на несколько частей. Графовые модели социальных сетей используются для моделирования экономических и коммуникационных связей людей, анализа процессов распространения информации, нахождения сообществ и связанных подгрупп, на которые можно разбить всю социальную сеть.

*Анализ центральности и других локальных свойств.* Чтобы определить относительную важность (вес) вершин графа (т. е. насколько участник в рамках конкретной сети является влиятельным), вводят понятие центральности – меры близости к центру графа. Центральность можно определить разными способами, поэтому существуют различные меры центральности. Следует отметить, что речь идет не о геометрической центральности при визуализации графа отношений.

*Центральность по степени* (*Degree centrality*) определяется как количество связей, инцидентных вершине:

Выделяют входящие и исходящие связи. Входящие связи характеризуют популярность человека, выходящие – его общительность. Полученную величину можно нормировать, разделив на общее число участников в сети.

Другими словами, центральность по степени предполагает, что среди участников сети более влиятельным является тот, у кого больше друзей, либо тот, кто входит в большее количество сообществ. Тем не менее участник сети, имеющий большое количество друзей, может быть связан с остальным графом маленьким количеством ребер. Поэтому вводится следующее понятие.

*Центральность по близости* (*Closeness centrality*) является показателем, насколько быстро распространяется информация в сети от одного участника к остальным. В качестве меры расстояния между двумя участниками используется кратчайший путь по графу (геодезическое расстояние). Так, непосредственные друзья участника находятся на расстоянии 1, друзья друзей – на расстоянии 2, друзья друзей друзей – на расстоянии 3 и т. д. Далее берется сумма всех расстояний и нормируется. Полученная величина называется удаленностью вершины

---

где кратчайший путь от вершины *v* до вершины *t*.

Другими словами, центральность по близости позволяет понять, насколько близок рассматриваемый участник ко всем остальным участникам сети. Таким образом, важно не только наличие непосредственных друзей, но и чтобы у самих этих друзей тоже были друзья.

*Центральность по посредничеству* (*Betweenness centrality*). Еще одной характеристикой участника является его важность при распространении информации. Именно в этом контексте центральность по посредничеству оценивает участника. Она рассчитывается как число кратчайших путей

между всеми парами участников, проходящих через рассматриваемого участника:

где

– общее количество кратчайших путей из вершины *s* к вершине *t*;

– количество кратчайших путей из вершины *s* к вершине *t*, проходящих через вершину *v*.

Для нормализации нужно разделить на количество пар вершин, за исключением самой вершины v, т.е. для ориентированного графа нужно разделить на , для неориентированного – на величину, равную /2. Недостатком центральности по посредничеству является ее вычислительная сложность.

Центральность по собственному вектору (Eigenvector centrality). Пусть центральность рассматриваемого участника – , а центральность его непосредственных друзей (соседних вершин) и т. д. Центральность по собственному вектору определяется как сумма центральностей соседних вершин, поделенных на константу λ, т. е.
. Выписав аналогичные уравнения для всех друзей, получим вектор неизвестных . Правила сложения определяются матрицей смежности т. е. = 1, если вершина v соединена с вершиной t, = 0 – иначе. Далее требуется решить уравнение AX = λX , т. е. найти собственные значения и собственные векторы матрицы A. Полученную задачу можно переписать иначе:

где – множество вершин, соседних с вершиной $v$; $\lambda$ – константа.

Собственный вектор, соответствующий самому большому собственному значению, как раз образован центральностями соответствующих участников сети. Таким образом, чем больше у участника друзей и чем они центральнее, тем больше его центральность. Верно и обратное: чем больше центральность участника, тем больше центральность его друзей. Недостатком центральности по собственному вектору также является вычислительная сложность.

Обобщением центральности по степени является *центральность Каца* (*Katz centrality*). Отличие в том, что центральность по степени учитывает количество непосредственных соседей вершины, а центральность Каца учитывает количество всех вершин, которые могут быть соединены путем

где $\alpha \in (0,1)$– доля участия удаленных вершин, называемая *коэффициентом затухания*.

Центральность Каца можно представить как разновидность центральности по собственному вектору:

Центральность можно вычислить при помощи *алгоритма ссылочного ранжирования* (*PageRank*), который используется в поисковой системе Google. В основу положен принцип «важности» веб-страницы: чем больше ссылок на страницу, тем она «важнее». Кроме того, вес самой страницы

определяется весом ссылки передаваемой на нее страницы. Таким образом, PageRank – это метод вычисления веса страницы путем подсчета важности ссылок на нее, т. е. вершина, ссылающаяся на другую вершину с большим весом, сама получает большой вес:

где – количество вершин, соседних с вершиной *j* (или количество выходящих связей в ориентированном графе).

Отличием данного алгоритма от вычисления центральности по собственному вектору и центральности Каца является наличие коэффициента пересчета . Следует заметить также, что в алгоритме ссылочного ранжирования используется обратная индексация матрицы смежности в сравнении с вычислением центральности по собственному вектору. Предшественником алгоритма PageRank является алгоритм *HITS* (*Hyperlink-Induced Topic Search*), предложенный Кляйнбергом [5].

Помимо перечисленных методов определения центральности, существует большое количество введенных неклассическим образом способов вычисления этой характеристики сети.

Важными характеристиками связей сети являются сбалансированность и транзитивность. Сбалансированность – это отсутствие ситуаций типа «позитивное взаимодействие (дружба, партнерство) между и , а также между и , но негативное взаимодействие (вражда, соперничество) между и ». Утверждается, что сбалансированные сети психологически более комфортабельны для участников и более устойчивы по сравнению с несбалансированными [6]. Транзитивность – это выполнение условий вида «если есть взаимодействие между и , а также между и , то имеет место взаимодействие между и ». Данные

характеристики описывают локальные связи участников и часто используются при анализе диад и триад.

Полезной характеристикой при анализе социальных сетей является уровень доверия. *Алгоритм вычисления уровня доверия* (*TrustRank)* предложен в [7]. Изначально был создан для отделения информативных веб-страниц от спама. Если говорить об этом алгоритме в терминах сайтов, для контрольной выборки эксперты вручную оценивают степень доверия небольшого количества сайтов, которые можно считать надежными. Эти сайты принимаются за эталон. Далее в основу алгоритма положено утверждение, что хорошие сайты редко ссылаются на плохие, а вот плохие очень часто ссылаются на хорошие. TrustRank – величина, которая дает оценку того, можно ли доверять конкретному сайту, считая, что он не содержит спама. Чем больше ссылок на сайте, тем меньше доверия «передается» по каждой такой ссылке. Степень доверия сайту (TrustRank) убывает с увеличением расстояния между ним и первоначальной выборкой. *Сила структурной позиции участника* является основным показателем, определяющим различия в ресурсах участников сети. В теории сетевого обмена для измерения данной характеристики вводится [8] индекс GPI силы участника (Genuine Progress Indicator):

где – число непересекающихся путей длины *k*, проходящих через вершину . Сила участника по сравнению с силой участника вычисляется как = .

*Методы обнаружения сообществ и анализ связанных подгрупп.* Связанные подгруппы (сообщества) в сети характеризуются наличием большого числа связей между входящими в них участниками и существенно меньшим числом

связей с остальными участниками. Анализ сообществ позволяет изучать устойчивость социальных структур. Простейший случай связанной группы – это сообщество, где каждый участник связан с каждым, и в данную группу не могут быть включены другие участники сети, поскольку они не имеют связей со всеми членами сообщества (клики). Таким образом, клика – это максимальный полный подграф данного графа. Если анализировать процессы распространения информации в графах, то можно дать другое определение сообщества, как множества участников, где путь между двумя любыми участниками не содержит более одной промежуточной вершины. В результате информация от одного участника к другому в связанной группе передается с минимальными искажениями. Связанные группы также могут быть выделены с помощью многомерного шкалирования или факторного анализа матрицы связей графа [1].

Для анализа устойчивости групповой структуры во времени используется следующая техника. Вначале строится трехмерная матрица, в которой строки представляют оценки взаимодействий участника со всеми другими участниками, данные самими участниками; столбцы являются собственными оценками взаимодействий участника; на третьей оси расположены периоды времени. Далее может быть построен график, показывающий изменения структуры подгрупп с течением времени.

После этого применяются *методы уменьшения размерности данных* (например, метод главных компонент), т. е. рассматривается проекция вершин сети в евклидово пространство пониженной размерности для описания зависимостей между строками и столбцами данной матрицы. В результате можно визуализировать изменения статуса пользователя сети на фоне изменений статусов подгрупп [6].

Полученную проекцию можно кластеризовать при помощи стандартных *алгоритмов кластеризации* как статистических (например, *k*-средних) [9], так и иерархических. Преимуществом иерархических методов является возможность

представления результата кластеризации в виде дендрограммы, т. е. на выходе будет получено не просто разбиение графа на группы, а иерархия групп и подгрупп в графе. Основная сложность подобных методов – подобрать подходящую меру расстояния (кратчайшего пути между вершинами) или меру сходства (similarity). Наиболее часто применяются меры сходства, использующие косинусный коэффициент (cosine similarity, также известен как коэффициент Охаи) и коэффициент Жаккара (Jaccard coefficient). Кластеризацию можно проводить не снизу вверх, а *сверху вниз*, т. е. сначала вся сеть рассматривается как одна группа, а на каждой итерации происходит последовательное отделение по одной связи. Детальный обзор методов обнаружения сообществ можно найти, например, в [10].

*Структурная эквивалентность участников сети.* Этот подход является противоположностью исследованию связанных групп. Участники эквивалентны, когда они занимают одинаковые позиции в социальной структуре сети, т. е. когда эквивалентны структура и тип взаимодействий этих участников с другими, при этом эквивалентные участники сети не должны взаимодействовать друг с другом.

В качестве меры эквивалентности может выступать плотность связей со структурными подгруппами участников сети [11]. Наряду со структурной эквивалентностью используется регулярная эквивалентность участников. В этом случае участники эквивалентны, когда они одинаковым образом взаимодействуют с участниками одного типа. Для определения структурной эквивалентности двух участников необходимо сравнить структуру их взаимодействий с другими участниками, т. е. нужно сравнить соответствующие столбцы в матрице связей графа. Это может быть осуществлено с помощью вычисления расстояния между этими векторами (например, по метрике Евклида или Чебышева) или коэффициентов связи (например, корреляции Пирсона). Для направленных графов необходимо учитывать входящие и

выходящие ребра, с этой целью одновременно рассматриваются две соответствующие матрицы.

На следующем этапе в матрицах для каждого типа связей переставляются столбцы таким образом, чтобы сгруппировать те из них, которые соответствуют структурно эквивалентным участникам. В результате матрица разбивается на структурные блоки, в каждом из которых вычисляется плотность. Далее строится новая матрица связей между найденными структурными блоками, например, по следующему правилу: если плотность связей между двумя блоками выше, чем средняя плотность связей в первоначальной матрице, то соответствующий элемент новой матрицы равен 1, в противоположном случае он равен 0. Такие матрицы называются блоковыми моделями и являются средством построения ролевых алгебр [1].

*Ролевые алгебры.* Это направление анализа социальных сетей фокусируется на выявлении логики взаимодействий участников сети в блоковых моделях, что позволяет выявлять сходства принципов взаимоотношений участников в различных социальных сетях. Определим, например, матрицы симпатии и антипатии следующим образом:

Теперь можем анализировать комбинации взаимодействий участников сети, перемножая соответствующие матрицы.

*Анализ диад и триад.* Диады – это набор из двух участников сети (вершин) и всех взаимодействий (ребер) между ними. Диада для каждого типа взаимодействий может находиться в одном из четырех состояний: нет связи между участниками, связь направлена от первого участника ко второму, связь направлена от второго участника к первому, взаимные связи участников. Анализ диад помогает установить вероятность наличия ребра между ними, степень зависимости

от свойств участников, определить условия и направления передачи информации и т.д. Для триад (три взаимодействующих участника) дополнительно исследуются вопросы транзитивности взаимодействий. Важным показателем является сила связей между участниками, которая определяется как линейная комбинация продолжительности, эмоциональной насыщенности, интимности или конфиденциальности и значимости взаимных услуг, которые характеризуют данное взаимодействие и соответствующее ему ребро графа. Слабые связи являются важными источниками информации [4], так как они помогают получить дополнительные сведения об участнике или сообществе, в котором он состоит, из других источников.

*Стохастические модели.* Основная идея вероятностных моделей ориентированных графов состоит в том, что каждая социальная сеть может быть рассмотрена как реализация случайного двумерного бинарного массива. Так как элементы этого массива являются зависимыми случайными величинами, то можно анализировать структуру зависимостей между соответствующими участниками социальной сети, находить вероятности существования определенных связей и получать оценки различных параметров сети.

Применение статистических моделей в анализе социальных сетей приведено, например, в [12]. Предлагается также применять методы машинного обучения и анализа данных для вычисления относительной автокорреляции, плотности связей и некоторых других характеристик сети. Более подробно про стохастические модели можно посмотреть в [1].

*Модели динамики сети.* Для определения динамики сети используются [13; 14] следующие модели.

*Модели эволюции графа.* Согласно этой модели, когда новая вершина добавляется к сети, происходит выбор вершин, к которым можно осуществить присоединение при помощи присоединяющего правила предпочтений. Также выбор вершины может осуществляться случайным образом или

«копированием» некоторых ее внешних ссылок. В [15] было эмпирически обнаружено, что сети со временем становятся плотнее (плотность сети увеличивается), т. е. количество ребер увеличивается линейно с ростом количества вершин. Более того, плотность сети меняется по степенному закону. В этой же работе описано еще одно наблюдение: диаметр сети часто уменьшается с течением времени, что противоречит общепринятому мнению о том, что меры расстояний должны медленно увеличиваться в зависимости от количества вершин.

*Модель* «*закрытого треугольника*» (Triangle-Closing Model) [15] утверждает, что новые вершины, добавленные к сети, имеют тенденцию к закрытию треугольника. Если считать, что связи, возникающие между участниками, образуют треугольник, то «открытый» треугольник возникает, когда два участника могут быть связаны друг с другом только посредст-вом третьего, т. е. одна из трех связей пропущена. Когда добавляется третья связь, получается «закрытый» треугольник.

*Модель* «*лесных пожаров*» [16] является в некотором смысле обобщением модели закрытого треугольника. Новая вершина присоединяется к существующей путем выбора подграфа, содержащего эту вершину и связывающего ее со всеми вершинами этого подграфа. Процесс начинается в выбранной вершине и напоминает распространение пожара через все вершины сети.

Несмотря на то, что существует довольно много работ, посвященных анализу глобальных свойств эволюции социальных сетей, вопросу эволюции графов на микроскопическом уровне посвящено совсем небольшое количество исследований. К этому направлению можно отнести, например, работы [14, 16], изучающие различные стратегии формирования сети и показывающие, что расположение ребер играет важную роль в эволюции сетей.

Среди работ, представляющих алгоритмические инструменты для анализа эволюции сетей, можно выделить [17], в которой предложены алгоритмы оценки

принадлежности пользователя сообществу и ее изменения с течением времени. Алгоритмы базируются на динамическом программировании, полном переборе, максимальном соответствии и «жадных» эвристиках. Основное внимание уделяется определению приблизительных кластеров пользователей и их временным изменениям. В [18] предложено применить принцип минимума длины описания MDL (Minimum Description Length) для поиска закономерностей в данных и обнаружения сообществ в динамических сетях, который создает структуру, «свободную от параметров». В [19] предлагается использовать принцип MDL для мониторинга эволюции сети.

*Анализ графов развития сети.* В работе [20] представлены различные подходы к анализу эволюции сети, основанные на парадигме извлечения ассоциативного правила (associationrule mining) и анализа частотной модели (frequent-pattern mining). Вводятся правила эволюции графа, новый тип частотных моделей, и рассматривается проблема поиска типичных моделей структурных изменений в динамических сетях. Сначала вычисляется набор частотных моделей графа, который описывает характерные эволюционные механизмы, а затем находят правила эволюции графа, которые удовлетворяют заданному ограничению минимальности доверия.

Проблема получения временно развивающихся веб-графов рассмотрена в [21].

Определены три уровня анализа графов: графы с единственной вершиной, подграфы и анализ графа в целом – для каждого из них используются свои техники. Изучены изменения свойств на каждом из трех уровней анализа. Для представления изменений в подграфах динамического графа в [22] предложен быстрый метод извлечения часто встречающихся «подпосле-довательностей» из графа. Однако в описании модели не оговаривается время, в течение которого наблюдались изменения графа во времени. Другой способ определения подграфов, меняющихся со временем, описан в

[23]. Он учитывает оценку важности вершин (vertex importance scores) и изменения близости вершин (vertex closeness changes). Релевантным подграфом считается не наиболее частый, а наиболее значимый.

Историю ребра в динамическом графе предложено в [24] представить в виде последовательности нулей и единиц, соответствующих наличию или отсутствию того или иного ребра. Затем для получения частотных моделей графа применяются традиционные методы получения графа. Разработанный алгоритм GREW использует эвристики и, вообще говоря, не извлекает все частые модели.

*Прогнозирование формирования связей.* Модели эволюции графа обычно создаются для оценки общестатистистических свойств существующих графов. Можно пытаться также вычислить, будут ли две конкретные вершины соединены друг с другом через некоторый промежуток времени. Это вычислительная задача, в основе которой лежит анализ эволюции социальной сети во времени, и называется проблемой прогнозирования связей.

Пусть дана краткая характеристика социальной сети в момент времени *t* и задано будущее время . Задача состоит в том, чтобы предсказать новые связи, которые, скорее всего, появятся в сети за промежуток времени Для ее решения в [25] применяется автоматическое моделирование процесса развития социальной сети с привлечением некоторых характеристик сети, таких как количество общих соседей, геодезическое расстояние (кратчайший путь), влиятельность вершины, момент первого попадания в социальную сеть.

Есть модели прогнозирования возникновения связей, основанные на машинном обучении, использующие личную информацию о пользователях сети для повышения точности предсказания. Иногда применяют иерархические, вероятностные (марковские) и реляционные модели для обнаружения связей между пользователями.

В других моделях [26] за основу предлагается брать сами свойства пользователей, и, например, наличие большого

количества связей (в блогосфере) может быть объяснено путем сопоставления демографических групп, общих интересов или географической близостью.

*Методы на основе онтологий.* Исследования [27] показали, что оценить параметры социальных сетей (диаметр, количество участников, среднюю длину пути и др.) можно при помощи онтологий. Сначала производится анализ видов элементов сети: люди, объекты (музыка, фото, видео, сообщение), взаимодействия (знает, сообщает, комментирует и т. д.). Затем авторы использовали существующие ресурсы онтологий и добавили варианты всевозможных связей, включая «папа», «мама», «друг», применили онтологию FOAF для определения участников социальной сети и контента, который они добавляют в сеть. Для описания тегов использовали новую версию SCOT. Была создана онтология SemSNI (Semantic Social Network Interactions) взаимодействий в социальной сети (посещений страниц, комментариев, личных сообщений) и онтология для анализа социальных сетей SemSNA. При помощи этих онтологий в рамках семантического анализа социальной сети удалось вычислить параметры подграфов социальной сети по раз-ным типам семантических связей («семья»/«family», «мне нравится» / «favorite», «дружба»/«isFriendOf») и типам взаимодействий («комментирует», «создает сообщение» и др.).

## 1.5. Программные приложения для анализа социальных сетей

Для анализа социальных сетей существует множество приложений для моделирования взаимодействий и процессов в сети, для вычисления определенных параметров сети и для визуализации графа сети. Например, приложения по визуализации сети ВКонтакте (см. http://www.yasiv.com/vk) или Facebook (http://www.touchgraph.com/facebook). В них используются различные методы и алгоритмы, которые описаны ранее в данной работе. К наиболее известным средствам

автома-тического анализа социальных взаимодействий относятся: NetMiner (http://www.netminer.com/ index.php), NetworkX (http:// networkx.lanl.gov), SNAP (http://snap.stanford.edu), UCINet (http://www.analytictech.com/ ucinet), Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek), ORA (см. http://www.casos.cs.cmu.edu/ тprojects/ora), Cytoscape (http://www.cytoscape.org) и др.

Для подобных приложений важным требованием является возможность обрабатывать очень большое количество данных. В связи с этим процесс обработки часто распараллеливают. Существуют приложения, которые моделируют «теорию шести рукопожатий», которые выстраивают цепочку из связей (друзей) между двумя пользователями сети: для русскоязычной сети ВКонтакте (http://ienot.ru/hand), для англоязычных сетей (http://www.sixdegrees.org, http://sixdegrees.com). Эти цепочки, как правило, действительно получаются небольшой длины.

Более подробную информацию о существующих приложениях для анализа социальных сетей можно найти, например, в [28; 29].

# 2. ЭМПИРИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ И МАТЕМАТИЧЕСКИЙ ФОРМАЛИЗМ

Так как большинство моделей информационного поиска учитывают статистическую природу текстов, остановимся на вопросах статистического распределения текстовых данных. В естественных науках хорошо известны и широко распространены такие статистические распределения, как Гауссово, показательное, биноминальное. Однако, в практике информационного поиска самое большое распространение имеет степенное распределение.

## 2.1. Эмпирические закономерности

Ниже будут обсуждаться параметры некоторых распределений, присущих многим информационным процессам, с учетом которых можно строить модели одновременно в рамках теории информационного поиска и концепции сложных сетей.

В 1965 году Д. Прайсом был впервые эмпирически обнаружен степенной закон распределения узлов по числу связей. Открытие в социальных сетях явления «тесного мира» С. Милграмом стало решающим фактором развития современной теории сложных сетей.

Важные явления и закономерности были обнаружены и исследованы в социальных сетях друзей и знакомств в последние годы. Оказалось, например, что в сетях друзей действует закон трех рубежей влияния: наше влияние распространяется только на наших друзей и друзей наших друзей. На следующем шаге это влияние уже ничтожно мало. Обратное также справедливо: наибольшее влияние на нас оказывают наши друзья и друзья наших друзей.

Еще на ранних этапах развития теории сложных сетей были детально исследованы законы распространения инфекционных заболеваний в социальных сетях, в том числе в сетях друзей и знакомых. Исследования последних лет

показали, что аналогичным образом распространяются в социальных сетях хорошее настроение (happiness) и депрессия, курение, алкоголизм, ожирениями- даже суицидальное поведение.

Когда при измерениях какой-либо величины вероятность получения того или иного значения обратна пропорциональна некоторой степени этого значения, говорят, что данная величина характеризуется степенным законом. Иногда также говорят о законе Зипфа или распределении Парето. Степенные законы часто встречаются в физике, биологии, науках о Земле и космосе, в экономике и финансах, информатике, демографии и прочих социальных науках.

В качестве основной функции, применяемой при статистическом методе описания, выступает *функция распределения*, которая определяет статистические характеристики рассматриваемой системы. Знание её изменения с течением времени позволяет описывать поведение системы со временем.

Дня введения понятия функции распределения сначала рассмотрим какую-либо макроскопическую систему, состояние которой описывается некоторым параметром *x,* принимающим *K* дискретных значений: Пусть при проведении над системой *N* измерений были получены следующие результаты: значение наблюдалось при измерениях, значение $x_2$ наблюдалось соответственно при измерениях и т.д. При этом, очевидно, что общее число измерений *N* равняется сумме всех измерений , в которых были получены значения .

Увеличение числа проведенных экспериментов до бесконечности приводит к стремлению отношения к пределу

Величина называется *вероятностью измерения значения* .

Вероятность представляет собой величину, которая может принимать значения в интервале 0 < < 1. Значение = 0 соответствует случаю, когда ни при одном измерении не наблюдается значение и, следовательно, система не может иметь состояние, характеризующееся параметром . Соответственно вероятность возможна только, если при всех измерениях наблюдалось только значение . В этом случае, система находится в детерминированном состоянии с параметром .

Сумма вероятностей нахождения системы во всех состояниях с параметрами равна единице:

---

Рассмотренный нами случай, когда параметр, характеризующий систему, принимает набор дискретных значений не является типичным при описании систем.

Рассмотрим статистическое описание, применимое для случая, когда измеренный параметр $x$ может иметь любые значения в некотором интервале . Причем, указанный интервал может быть и не ограниченным какими либо конечными значениями и . В частности параметр $x$ в принципе может изменяться от до + .

Пусть в результате измерений было установлено, что величина $x$ с вероятностью попадает в интервал значений от $x$ до $x + dx$. Тогда можно ввести функцию , характеризующую плотность распределения вероятностей (probability density function (PDF)):

---

Эта функция в физике обычно называется *функцией распределения.*

Функция распределения должна удовлетворять условию: , так как вероятность попадания измеренного значения в интервал от $x$ до $x + dx$ не может быть отрицательной величиной. Вероятность того, что измеренное значение попадет в интервал $< x <$ равна

Соответственно, вероятность попадания измеренного значения в весь интервал возможных значений равна единице:

(2.1)

Выражение (2.1) называется *условием нормировки функции распределения.*

Когда при измерениях какой-либо величины вероятность получения того или иного значения обратна пропорциональна некоторой степени этого значения, говорят, что данная величина характеризуется степенным законом. Иногда также говорят о законе Зипфа или распределении Парето.

Рассмотрим кумулятивные распределения социальных систем различного рода. Все они следуют степенному распределению по меньшей мере в некоторой части диапазона.

Что впервые было замечено Прайсом [30], число цитирований, сделанных на те или иные научные статьи, соответствует степенному распределению (рис. 2.1). Данные взяты из «Индекса научного цитирования», составленного Реднером [31] для статей, опубликованных в 1981 году. Диаграмма представляет кумулятивное распределение числа

цитирований статей, сделанных с момента публикации статей до июня 1997 года.

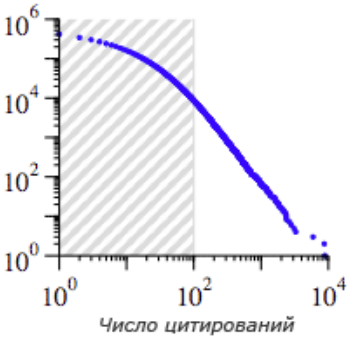



Рис. 2.1. Число цитирований научных статей, которые были опубликованы в 1981 году, сделанных с момента публикации до июня 1997 года

Кумулятивное распределение числа «хитов», полученных веб-сайтами (то есть, серверами, а не отдельными веб-страницами) в течение одного дня, совершенных некоторым подмножеством пользователей интернет-провайдера AOL. (рис. 2.2). Сайтом, получившим с большим отрывом больше всего хитов является yahoo.com. Данные исследования Адамика и Губермана [30].

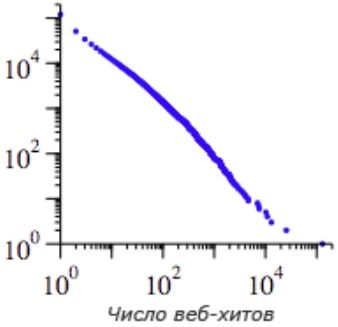



Рис. 2.2. Число хитов 60 тыс. пользователей интеренет-провайдера AOL, полученных веб-сайтами в течение одного дня 1 декабря 1997 года

Как мы видим, степенные распределения удивительно широко распространены, однако они являются не единственной формой «широких» распределений. Поскольку могло сложиться впечатление, что все интересные распределения являются сте-пенными, позвольте подчеркнуть, что существует немало нату-ральных величин, обладающих скошенными влево распре-делениями, которые тем не менее не являются степенными распределениями. Вот несколько примеров (рис. 2.3):

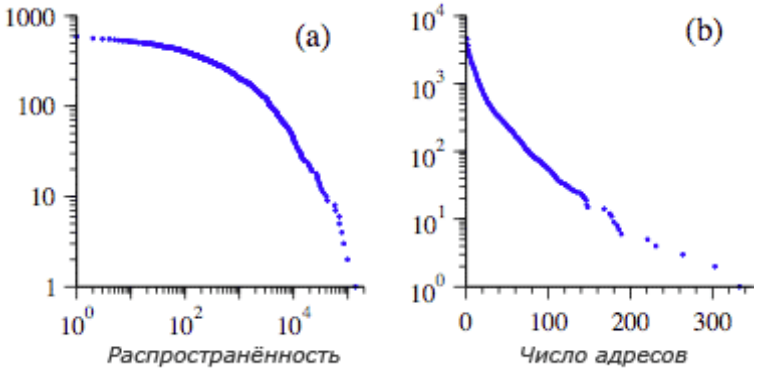



Рис. 2.3. Кумулятивные распределения некоторых величин, распределения которых охватывают несколько порядков, но при этом не соответствуют степенному закону. (a) Число наблюдений североамериканских птиц, принадлежащих 591 виду по данным North American Breeding Bird Survey 2003. (b) Число записей в адресных книгах электронной почты 16881 пользователей большой университетской компьютерной системы [32]

### 2.1.1. Распределение Парето

Анализируя общественные процессы, В. Парето (V. Pareto) рассмотрел социальную среду как пирамиду, на вершине которой находятся люди, представляющие элиту.

Парето в 1906 году установил, что около 80 процентов земли в Италии принадлежит лишь 20 процентам ее жителей. Он пришел к заключению, что параметры полученного им распределения приблизительно одинаковы и принципиально не различаются в разных странах и в разное время. Парето также установил, что точно такая же закономерность наблюдается и в распределении доходов между людьми, которое описывается уравнением , где – величина дохода, - количество людей с доходом, равным или превышающим , и - параметры распределения. В математической статистике это распределение получило имя Парето, при этом предполагаются естественные ограничения на параметры: . Распределению Парето присуще свойство устойчивости, т.е. сумма двух случайных переменных, которые имеют распределение Парето, также будет распределена по Парето. Замеченное распределение, называемое «законом Парето» или «принципом 80/20», применимо в очень многих областях. Например, при информационном поиске достаточно определить 20% важнейших ключевых слов, чтобы найти 80% необходимых документов, а затем расширить поиск или воспользоваться опцией "найти похожие" для полного решения задачи. Еще один пример: 80% посещений веб-сайта приходится лишь на 20% его веб-страниц.

При построении систем массового обслуживания, в том числе и информационно-поисковых систем, необходимо учитывать тот факт, что наиболее сложным функциональным возможностям системы, на реализацию которых уходит 80 и больше процентов трудозатрат, будут пользоваться не более чем 20 процентов пользователей данной системы.

Перейдем к более строгой формулировке закона Парето. Предположим, что последовательность соответствует размерам доходов отдельных людей. После ранжирования этой последовательности по убыванию получается новая последовательность (элементы расположены в порядке убывания).

Предположим, что – общее число людей, у которых доход составляет не менее , т.е. . Тогда правило Парето можно переписать в таком виде:

Отсюда:

Рассматривается сумма первых значений величины , то есть общая величина дохода наиболее богатых людей – составляет:

где

Переходя от дискретных величин к непрерывным (предполагая, что ), имеем:

В безразмерных переменных - и последнее равенство имеет вид (см. рис. 2.4):

Величина в нашем примере – относительное количество дохода, получаемого первыми по рангу людьми, доля которых (относительно всех людей) равна .

Для последних двух случаев, представленных на рис. 2.4, - 20% людей имеют - 80% доходов (близкие к этим значениям явления наблюдаются в реальной жизни).

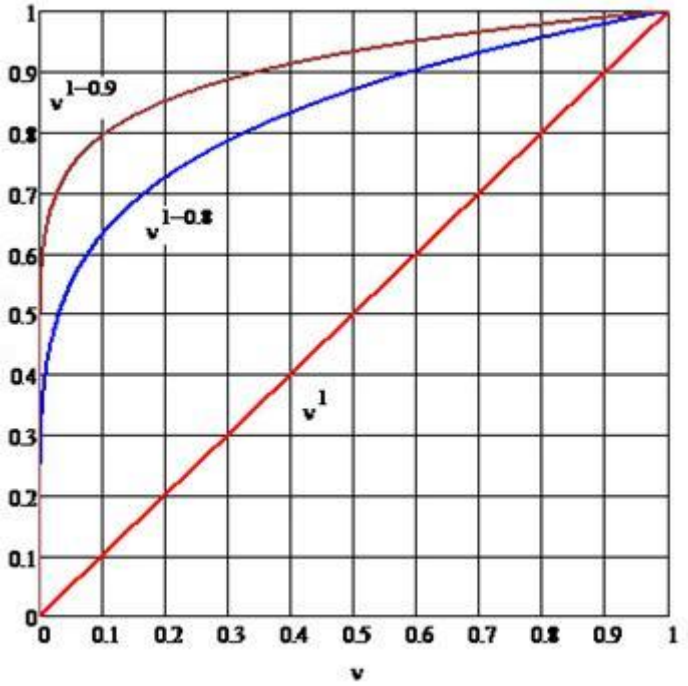



Рис. 2.4. Распределение Парето для различных значений параметров: зависимость для трех случаев – (доходы всех одинаковы) и

### 2.1.2. Законы Ципфа

Дж. Ципф (G. Zipf) изучал использование статистических свойств языка в текстовых документах и выявил несколько эмпирических законов, которые представил как эмпирическое доказательство своего «принципа наименьшего количества усилий». Он экспериментально показал, что распределение слов естественного языка подчиняется закону, который часто называют первым законом

Ципфа, относящимся к распределению частоты слов в тексте. Этот закон можно сформулировать таким образом. Если для какого-нибудь довольно большого текста составить список всех слов, которые встретились в нем, а потом ранжировать эти слова в порядке убывания частоты их появления в тексте, то для любого слова произведение его ранга и частоты появления будет величиной постоянной: , где – частота встречаемости слова в тексте; – ранг слова в списке; – эмпирическая постоянная величина (коэффициент Ципфа). Для славянских языков, в частности, коэффициент Ципфа составляет приблизительно 0,06-0,07.

Приведенная зависимость отражает тот факт, что существует небольшой словарь, который составляет большую часть слов текста. Это главным образом служебные слова. Например, приведенный в [42] анализ романа «Том Сойер», позволил выделить 11.000 английских слов. При этом было обнаружено двенадцать слов (the, and, и др.), каждое из которых охватывает более 1 % лексем в романе. Закон Ципфа был многократно проверен на многих массивах. Ципф объяснял приведенное выше гиперболическое распределение «принципом наименьшего количества усилий» предполагая, что при создании текста меньше усилий уходит на повторение некоторых слов, чем на использование новых, т.е. на обращение к «оперативной памяти, а не к долговременной».

Ципф сформулировал еще одну закономерность, так называемый второй закон Ципфа, состоящий в том, что частота и количество слов, которые входят в текст с данной частотой, также связанны подобным соотношением, а именно:

—

где – количество различных слов, каждое из которых используется в тексте раз;

– константа нормирования.

Существует простая количественная модель определения зависимости частоты от ранга. Предположим, что генерируется случайный текст обезьяной на пишущей машинке. С вероятностью генерируется пробел, а с вероятностью – другие символы, каждый из которых имеет равную вероятность. Показано, что полученный таким образом текст будет давать результаты, близкие по форме к распределению Ципфа. Эта модель была усовершенствована в соответствии с фактическими эмпирическими данными, когда вероятности генерации отдельных символов были заданы на основе анализа большого текстового массива [43]. Полученное соответствие не доказывает закона Ципфа, но вполне его объясняет с помощью простой модели.

Более сложную модель генерации случайного текста, удовлетворяющего второму закону Ципфа, предложил Г. А. Саймон (H. A. Simon). Условия этой модели достаточно просты: если текст достиг размера в слов, тогда то, каким будет -е слово текста определяется двумя допущениями:

1. Пусть – количество разных слов, каждое из которых использовалось раз среди первых слов текста. Тогда вероятность того, что -ым окажется слово, которое до того использовалось раз пропорционально - общему количеству появления всех слов, каждое из которых до этого использовалось раз.

2. С вероятностью -ым словом будет новое слово.

Распределение Ципфа часто искажается на практике ввиду недостаточных объемов текстовых корпусов, что приводит к проблеме оценки параметров статистических моделей. Вместе с тем соотношение между рангом и частотой была взята Солтоном в 1975 г. [44] как отправная точка для выбора терминов для индексирования. Далее им рассматривалась идея сортировки слов в соответствии с их частотой в текстовом массиве. Как второй шаг высокочастотные слова могут быть устранены, потому что они

не являются хорошими различительными признаками для отдельных документов из текстового массива. На третьем шаге термы с низкой частотой, определяемой некоторым порогом (например слова, которые встречаются только единожды или дважды) удаляются, потому что они встречаются так нечасто, что редко используются в запросах пользователей. Используя этот подход, можно значительно уменьшить размер индекса поисковой системы. Более принципиальный подход к подбору индексных термов – учет их весовых значений. В весовых моделях среднечастотные термы оказываются самыми весомыми, так как они являются наиболее существенными при отборе того или иного документа (наиболее частотные слова встречаются одновременно в большом количестве документов, а низкочастотные могут не входить в документы, интересующие пользователя).

Еще один эмпирический закон, сформулированный Ципфом состоит в том, что количество значений слова коррелирует с квадратным корнем его частоты. Подразумевалось, что нечасто используемые слова более однозначны, а это подтверждает то, что высокочастотные слова не подходят для внесения в индексы информационно-поисковых систем.

Ципф также определил, что длина слова обратно пропорциональна его частоте, что может быть легко проверено путем простого анализа списка служебных слов. Последний закон действительно служит примером принципа экономии усилий: более короткие слова требуют меньше усилий при воспроизведении, и таким образом, используются более часто. Этот «закон» можно подтвердить, рассматривая приведенную выше модель генерации слов обезьяной. Легко видеть, что вероятность генерации слова уменьшается с длиной, вероятность слова из  непробельных символов равна:

где  – вероятность генерации пробела.

Хотя закон Ципфа дает интересные общие характеристики слов в текстовых массивах, в общем случае замечены некоторые ограничения его применимости при получении статистических характеристик документальных массивов, состоящих из множества независимых документов разных авторов.

Законам Ципфа удовлетворяют не только слова из одного текста, но многие объекты современного информационного пространства.

### 2.1.3. Закономерность Бредфорда

Закономерность С. Бредфорда (S. Bredford), известного документалиста, одного из авторов универсальной десятичной классификации – УДК, состоит в следующем: если научные журналы расположить в порядке убывания числа помещенных в них статей по конкретному предмету, то полученный список можно разбить на три зоны таким образом, чтобы количество статей в каждой зоне по заданному предмету была одинаковой. Эти три зоны представляют: ядро - профильные журналы, непосредственно посвященные рассмотренной тематике, журналы, частично посвященные заданной области и журналы, тематика которых довольно далека от рассмотренного предмета. С. Бредфорд в 1934 г. установил следующее соотношение для количества журналов в разных зонах [45]:

— —

где количество журналов в первой зоне – , во второй – , в третьей – .

Бредфорд вначале рассматривал найденную закономерность только как специфический случай распределения Ципфа для системы периодических изданий по науке и технике. Однако в дальнейшем оказалось, что эта же закономерность справедлива и для периодических изданий из

многих других предметных областей [46, 47], а также для наборов веб-сайтов, относящихся к некоторой выбранной тематике.

### 2.1.4. Закон Хипса

В компьютерной лингвистике эмпирический закон Г.С. Хипса (H.S. Heaps) связывает объем документа с объемом словаря уникальных слов. Казалось бы, словарь уникальных слов должен насыщаться, а его объем стабилизироваться при увеличении объемов текста. Оказывается, это не так! Для всех известных сегодня текстов в соответствии с законом Хипса, эти значения связаны соотношением (рис. 2.5):

где -это объем словаря уникальных слов, составленный из текста, который состоит из уникальных слов, и – определенные эмпирически параметры. Для европейских языков принимает значение от 10 до 100, а - от 0.4 до 0.6.

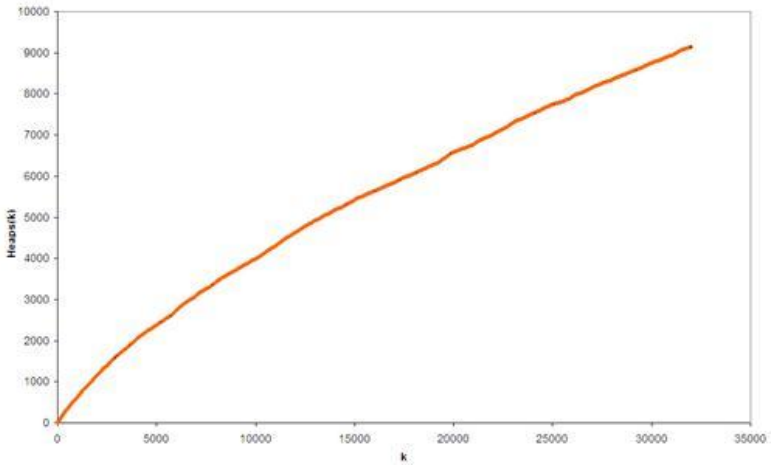



Рис. 2.5. Типичный график, подтверждающий закон Хипса: по оси абсцисс – количество слов в тексте, по оси ординат – объем словаря – количество уникальных слов

Закон Хипса справедлив не только для уникальных слов, но и для многих других информационных объектов, что вполне естественно, так как уже доказано [48], что он является следствием закона Ципфа.

## 2.2. Степенные распределения случайных величин

Наиболее частыми (как обычно считается), универсальными законами распределения случайных величин, встречаемыми в различных естественнонаучных исследованиях, является нормальный закон – распределение Гаусса и так называемое логнормальное распределение (рис. 2.6):

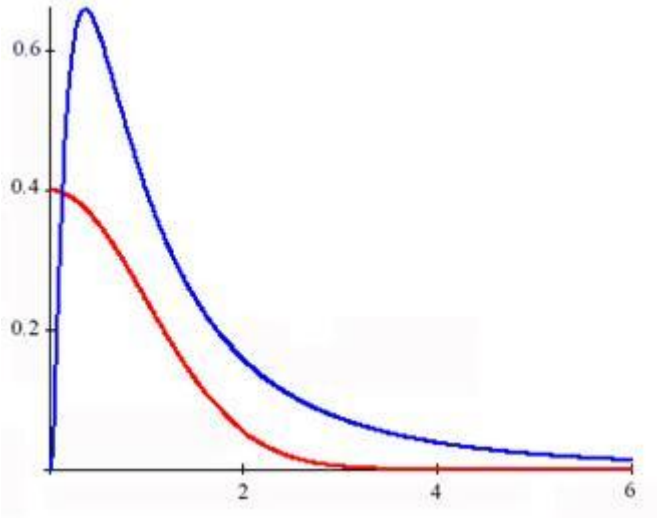



Рис. 2.6. Графики нормального и логнормального распределения. Среднее значение для нормального распределения выбрано равным нулю

Частая встречаемость нормального закона объясняется тем, что когда случайная величина является суммой независимых случайных величин, то ее распределение приближается к нормальному. Именно это утверждение является содержанием так называемой центральной предельной теоремы теории вероятностей. Заметим, что часто в конкретных исследованиях гауссово распределение случайной величины принимается в силу привычки или удобства.

Б. Мандельброт был одним из первых, кто обратил пристальное внимание на то, что не менее универсальным, часто встречаемым законом распределения случайной величины является степенное (часто говорят гиперболическое) распределение с плотностью вероятности:

—

или

—

где - вероятность того, что , а и - некоторые положительные константы, параметры распределения.

Следует отметить, что приведенное выше распределение рассматривалось Б. Мандельбротом (B. Mandelbrot) как уточнение закона Ципфа и его часто называют распределением Ципфа-Мандельброта. При этом оказалось, что – близкая к единице величина, которая может изменяться в зависимости от свойств текста и языка. Соответственно,

— ——

Справедливости ради надо отметить, что степенные функции распределения рассматривались еще Коши. Как

наглядный пример распределения Коши можно привести модель стрельбы из вращающегося в горизонтальной плоскости пулемета (рис. 2.7).

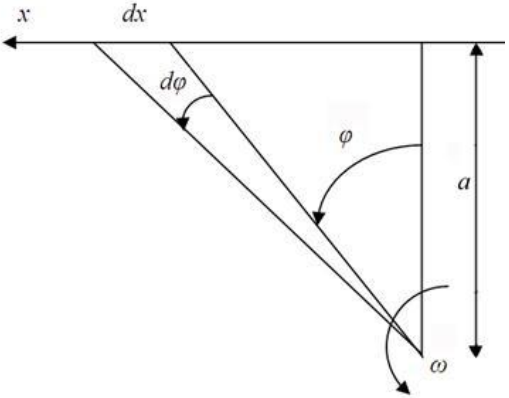



Рис. 2.7. Модель, приводящая к распределению Коши

Если, производя одиночные выстрелы, нажимать на курок равновероятно при любом его положении, то функция распределения выстрелов по углу – будет величиной постоянной: . С другой стороны, вероятность попадания в бесконечно малый участок бесконечной плоской мишени равна Откуда, с учетом после элементарных преобразований находим распределение Коши:

Так как для этой функции интеграл

не определён для , то ни математическое ожидание, ни дисперсия, ни моменты старших порядков этого распределения

не определены. В этом случае говорят, что математическое ожидание не определено, а дисперсия бесконечна.

Напомним, гиперболическое распределение названо в честь В. Парето, а дискретный закон распределения с ранжированной переменной был назван в честь Д. Ципфа, который сформулировал его для описания частоты употребления слов.

## 2.3. Масштабно-инвариантные распределения

Степенные распределения иногда именуются *масштабно-инвариантными распределениями*. Почему? Потому что степенные распределения – это единственный тип распределений, которые выглядят одинаково *вне зависимости от того, в каком масштабе мы их рассматриваем*. Разберемся, что это значит.

Предположим, мы имеем какое-то вероятностное распределение величины $x$, обозначенное как Пусть также мы обнаружили, что оно обладает следующим свойством:

(2.2)

для произвольного $b$. То есть, если мы увеличим шкалу или единицу измерений величины $x$ в $b$ раз, форма распределения не изменится, не считая общего множителя-константы.

Подобной масштабной инвариантностью не обладает большинство других распределений. Например, таким свойством не обладает экспоненциальное распределение. Более того, как мы сейчас увидим, единственный тип распределений, обладающих этим свойством - это степенные распределения.

Оттолкнемся от уравнения (2.2). При мы получим То есть, и мы можем записать уравнение (2.1) как

Поскольку мы полагаем, что это равенство должно быть верным при любом значении $b$, мы можем продифференцировать обе его части по $b$, и получим:

тут обозначает производную $p$ по аргументу $b$. Теперь положим

Это простое дифференциальное уравнение первого порядка, которое имеет решение:

Устанавливая $x=1$, мы выясним, что константа равна Теперь берём экспоненту обеих частей равенства:

где . Таким образом, как мы и предполагали, степенное распределение – единственный тип распределений, удовлетворяющих требованию масштабной инвариантности (2.2).

Существуют системы, которые приобретают масштабно-инвариантные свойства при определенных особых значениях управляющих параметров. Точки, которым соответствуют эти особые значения именуются «непрерывными фазовыми переходами», и судя по проведённому выше анализу, в таких

точках наблюдаемая статистика системы должна приходить к степенному распределению.

## 2.4. Степенные законы для дискретных переменных

До сих пор мы фокусировались на степенных распределениях непрерывных переменных, но многие величины, с которыми мы имеем дело на практике являются на самом деле дискретными – обычно, целочисленными. Например, социальная сеть, у сети есть узлы, а у узлов есть соседи (степени узлов). В данном случае это дискретные целочисленные значения. Во многих случаях эта особенность не очень важна. Часто степенному закону подчиняется только хвост распределения, где величины настолько велики, что во всех практических контекстах такие распределения можно считать непрерывными. Однако, технически, для целочисленных дискретных величин степенные распределения должны определяться несколько иначе.

Пусть $k$ – целочисленная переменная. Тогда один из возможных способов – заявить, что эта переменная соответствует степенному распределению, если вероятность появления значения $k$ в наборе замеров равна

(2.3)

при каком-то постоянном показателе $\alpha$. Ясно, что это распределение не может охватывать все значения вплоть до поскольку оно расходится в этом пределе, но теоретически распределение может продолжаться до Если мы отбросим все данные, для которых константа $C$ может быть получена из условия нормировки:

Тут - дзета-функция Римана. Преобразуя, мы обнаружим, что и

Если, как это часто бывает, степенной закон охватывает только хвост распределения, для значений выражение принимает вид:

где

это обобщенная или неполная дзета-функция.

Большинство результатов, полученных в предыдущих параграфах для непрерывных распределений можно обобщить и для дискретных, хотя математика оказывается сложнее и часто опирается на специальные функции вместо легко берущихся интегралов как в непрерывном случае.

Было выдвинуто соображение, что уравнение (2.3) – не самое лучшее представление степенного закона для дискретных величин. Альтернативная и во многих случаях более удобная форма:

где – это уже упоминавшаяся нами бета-функция Лежандра, Бета-функция ведет себя как степенная функция, то есть, при больших $k$, так что асимптотически распределение приобретает нужную степенную форму.

Распределение Юла удобно, поскольку суммы, которые получаются при расчете вероятностей на его основе часто принимают закрытый вид, в то время как суммы, могут быть записаны только с использованием специальных функций. Например, константа нормировки $C$ для распределения Юла задается уравнением:

$$\_\_\_\_$$

Отсюда , и

Первый и второй моменты (то есть, среднее и средний квадрат распределения):

$$\_\_\_\_ \qquad \_\_\_\_\_$$

Такие же простые выражения получаются и для многих других результатов, которые мы получали для непрерывного случая.

# 3. МОДЕЛИ ФОРМИРОВАНИЯ И РОСТА СЕТЕЙ

Теория случайных графов стала интенсивно развиваться с конца 50-х годов прошлого века после публикации статей Эрдеша-Реньи об эволюции случайных графов. В этой модели все ребра появляются случайно и независимо с одинаковой вероятностью  и под эволюцией понимается изменение свойств графов с ростом вероятности  . Оказалось, что в некоторых значениях $p$ происходит так называемый фазовый переход и свойства графа кардинально меняются. В этом направлении было получено много интересных и глубоких результатов. Однако, в начале 2000-х выяснилось, что модель Эрдеша-Реньи плохо описывает реальные графы, возникающие в различных областях, в частности в графы таких социальных сетей как Facebook, Twitter и т.п.

Это породило много новых исследований математических моделей случайных графов. В задачах описания динамиики социальных сетей основное значение имеет правильный выбор математической модели. На данный момент известно множество моделей случайных графов и безмасштабных (scale-free) сетей, некоторые из которых показали удовлетворительные результаты при сравнении с экспериментальными данными.

Вообще говоря, модели социальных сетей можно разделить на три класса: модели случайных графов (модель Эрдеша-Реньи и ее обобщения), простейшие модели безмасштабных сетей (модель Боллобаша и ее обобщения, модель копирования и др.) и более гибкие модели безмасштабных сетей (модель Чунг-Лу, модель Янсона-Лучака и др.) На наш взгляд третий класс моделей представляет наибольший интерес при моделировании больших реальных социальных сетей, таких как Facebook.

## 3.1. Модель Эрдеша-Реньи

Наиболее отвечающими действительности сетевыми моделями в настоящее время являются модели Эрдёша-Реньи и Барабаши-Альберт.

Модель Эрдёша-Реньи (Erdos-Renyi) определяют как случайный граф, имеющий N помеченных вершин, соединенных n ребрами, которые случайно выбираются из всех возможных ребер (рис. 3.1).

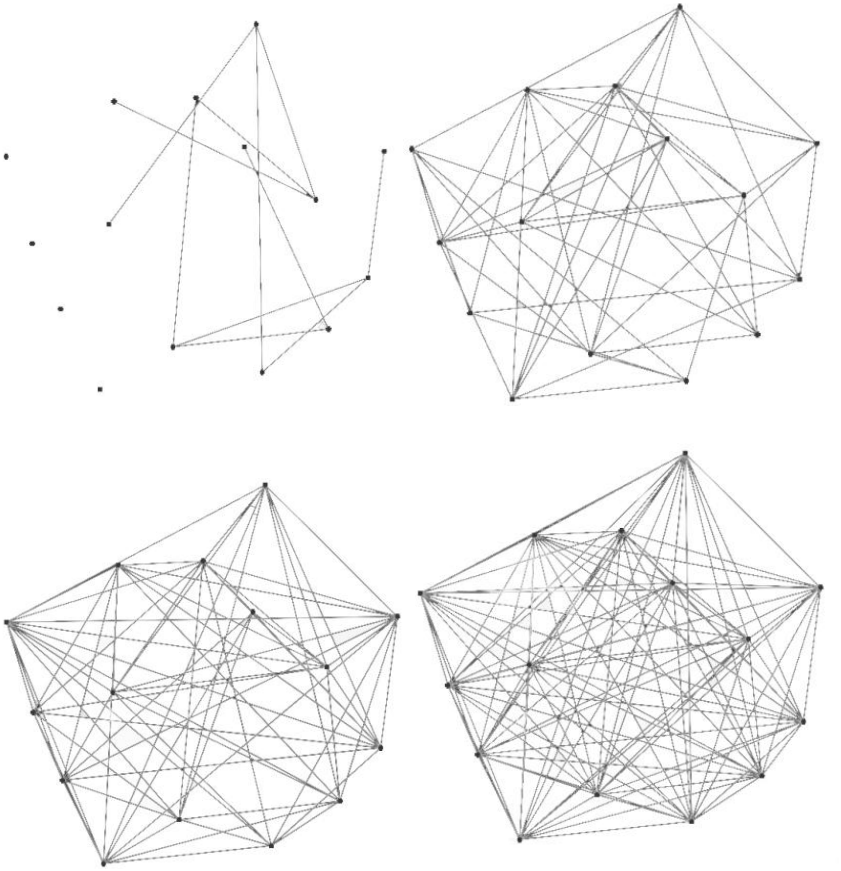


Рис. 3.1. Вид графа по модели Эрдёша-Реньи, состоящий из 15 вершин с разной вероятностью нахождения связей

Зафиксируем натуральное число и рассмотрим множество

Таким образом мы задали множество вершин случайного графа. Зададим полный граф на множестве вершин . Пронумеруем ребра , где

Зададим некоторое и будем выбирать ребра из множества согласно схеме Бернулли. Таким образом мы получили случайный граф Формально выражаясь, мы имеем вероятностное пространство в котором:

(3.1)

Таким образом, в модели Эрдеша-Реньи каждое ребро независимо от других ребер входит в случайный граф с вероятностью . Модель Эрдеша-Реньи на данный момент является самой изучаемой моделью случайных графов. Приведем несколько фактов о ней.

Будем далее говорить, что случайный граф обладает некоторым свойством почти наверное, если вероятность обладания этим свойством стремится к единице

**Треугольники в случайном графе**. Обозначим через случайную величину на пространстве , равную количеству треугольников в случайном графе . Тогда верны следующие три теоремы:

**Теорема 1** Пусть при Если — то почти наверное (т.е. граф не содержит треугольников).

**Теорема 2** Пусть – где константа. Тогда имеет асимптотически пуассоновское распределение с параметром —

**Теорема 3** Пусть при Тогда если

—— то

**Связность случайного графа.** Одно из самых интересных свойств модели Эрдеша-Реньи – наличие фазового перехода:

**Теорема 4** Пусть —— Если , то почти наверное случайный граф связен. Если , то почти наверное случайный граф связным не является.

**Теорема 5** Пусть = – Тогда при любом существует такая константа что почти наверное каждая компонента случайного графа имеет не более вершин. При любом существует такая константа , что почти наверное среди компонент случайного графа есть одна (гигантская), число вершин которой не меньше .

Таким образом, модель Эрдеша-Реньи и ее простейшие обобщения являются слишком негибкими для моделирования больших социальных сетей, т.к., например, в них нет соответствующих фазовых переходов и распределения числа треугольников.

## 3.2. Наблюдения Барабаши-Альберт

Сети со степенным распределением (3.2) степеней вершин называются безмасштабными (Scale-Free) [49]:

. (3.2)

Данная модель также называется моделью Барабаши-Альберт.

Барабаши и Альберт предложили простую и элегантную модель возникновения и эволюции безмасштабных сетей. Они показали, что для возникновения безмасштабных сетей необходимы два условия:

1. *Рост.* Начиная с небольшого числа $m_0$ узлов, на

каждом временном шаге добавляется один новый узел с m ($m \leq m_0$) связями, которые соединяют этот новый узел с m различными уже существующими узлами.

2. *Предпочтительное присоединение (Preferencial attachment)*. Когда выбираются узлы, к которым присоединяется новый узел, предполагается, что вероятность P с которой новой узел будет соединяться с уже существующим узлом, *i* зависит от числа связей,  , которыми этот узел уже связан с другими узлами, так что:

(3.3)

При степенном распределении возможно существование вершин с очень высокой степенью (рис. 3.2), что практически не наблюдается в сетях с распределением Пуассона. Данное свойство модели Барабаши-Альберт отвечает представлению реальных сетей, содержащих узлы с большим количеством связей (концентраторы, хабы и т.д.) (рис. 3.3).

Модель Барабаши-Альберт является более оптимальной по сравнению с моделью Эрдёша-Реньи с точки зрения удовлетворения типичных свойств реальных сетей, таких как коэффициент кластеризации, длина пути, корреляция степеней вершин [49].

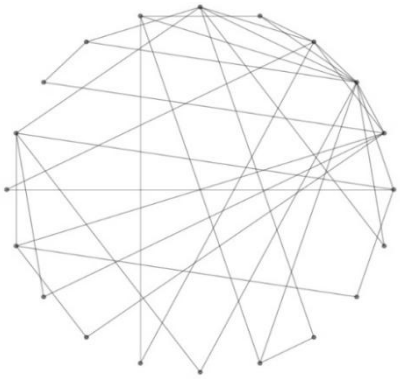


Рис. 3.2. Вид графа по модели Барабаши-Альберт, состоящий из 20 вершин

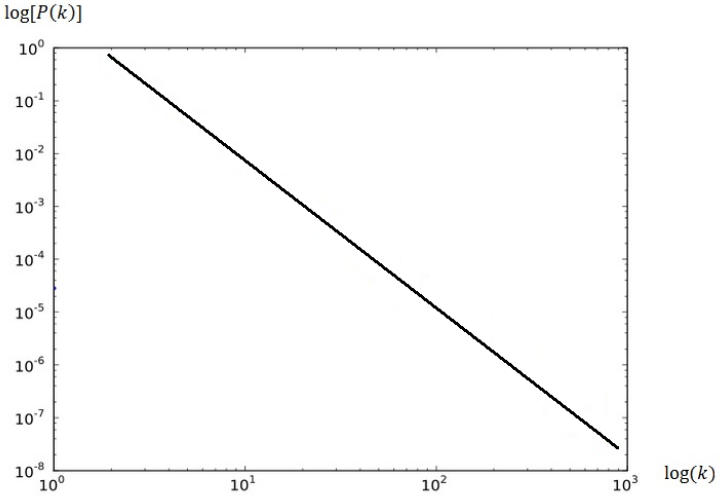



Рис. 3.3. Распределение степеней вершин модели Барабаши-Альберт в логарифмическом масштабе

На основании своих наблюдений авторы ввели понятие предпочтительного присоединения (preferential attachment). Рассмотрим процесс генерации графа. На *n*-ом шагу мы добавляем новую вершину *n* с *m* ребрами, инцидентыми ей, причем вероятность ребра к вершине i пропорциональна степени вершины i (см. рис. 3.4):

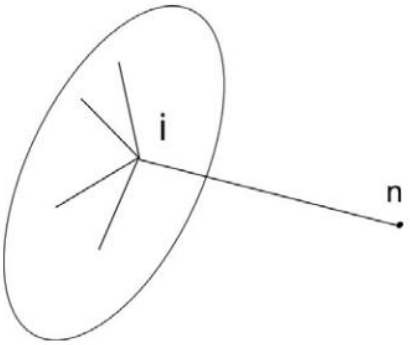



Рис. 3.4. Добавление вершины на n-ом шагу

——— (3.4)

Основных проблем со спецификацией модели Барабаши-Альберт две.

Во-первых, результирующий граф зависит от начального параметра *m*. Например, при модель Барабаши-Альберт описывает генрацию дерева, если начальный граф– тоже дерево. Если начальный граф несвязный, то и все последующие тоже будут таковыми.

Во-вторых, трудность с предпочительным присоединением заключается в случайном выборе вершин (если ), к которым присоединится новая вершина. Например, верна следующая теорема.

**Теорема 1** Пусть – произвольная целочисленная функция, такая что:

для любых и при Тогда существует такой процесс генерации случайного графа удовлетворяющий (1), что с вероятностью 1 в ровно треугольников для достаточно больших

Говоря менее формально, теорема 1 говорит о том, что если вы хотите иметь в графе с *n* вершинами log*n* треугольников, есть модель Барабаши-Альберт, которая выдаст такой результат.

——————— ~——————— (3.5)

где – количество ребер, имеющих вершину *i* своим левым концом в графе

Т.е. мы получили степенной закон От условия смогли избавиться сравнительно недавно, а для того чтобы получить степень 2.1 (соответствующую реальной

степени веб-графа несколько лет назад) вместо 3, надо отойти от модели Боллобаша-Риордана.

Пусть $H$– фиксированный граф. Обозначим через $\#(H,$ ) случайную величину, равную количеству подграфов графа изоморфных графу $H$.

**Теорема 2** Пусть Пусть также полный граф на трех вершинах. Тогда

(3.6)

при .

**Теорема 3** Пусть фиксированы и Пусть также цикл на $l$ вершинах. Тогда

(3.7)

при , где положительная константа. Более того, при имеем

А. Рябченко и Е. Самосват из Яндекса в модели, близкой к модели Боллобаша-Риордана, установили следующий факт.

**Теорема 4** Пусть задан граф $H$, степени вершин котором равны Обозначим через число вершин в $H$, степень каждой из которых равна $m$. Тогда

(3.8)

Зависимость от $k$ занесена в константу
По(6)

(3.9)

что согласуется с теоремой 2. А для теорема 4 говорит, что его средняя частота в веб-графе постоянна. Т.о., «тетраэдров» в веб-графе почти нет.

Следует отметить, что последнее утверждение имеет мало общего со свойствами реального веба: в нем встречаются и тетраэдры, и клики большей мощности. Это связано с действием спамеров и агентств по раскрутке сайтов (групп в социальных сетях). Спам в модели Боллобаша-Риордана не учитывается.

### 3.3. Модель LCD

Выделим в пространстве ось абсцисс и зафиксируем на ней *2n* точек: . Разобьем эти точки на пары, и элементы каждой пары соединим дугой, лежащей в верхней полуплоскости. Полученный объект назовем линейной хордовой диаграммой (LCD). Дуги в LCD могут как пересекаться, так и лежать друг под другом, но не могут иметь общих вершин. Количество различных диаграмм равно

— (3.10)

По каждой диаграмме построим граф с *n* вершинами и *n* ребрами. Процесс построения описан в алгоритме1 и показан на рис. 3.5.

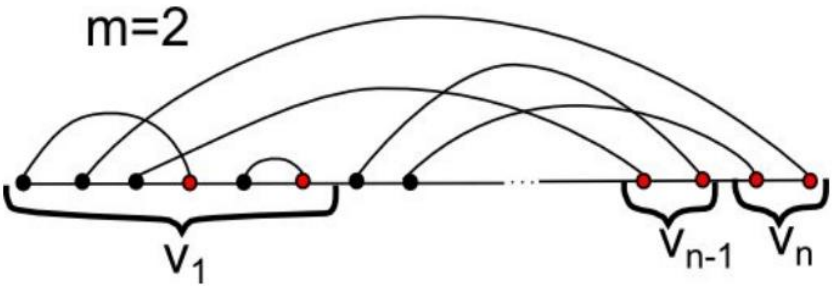



Рис. 3.5. LCD модель

**Алгоритм1** Алгоритм построения графа по LCD:

1) Идем слева направо по оси абсцисс, пока не встретим правый конец какой либо дуги.
2) Пусть этот конец имеет номер .
3) Обьявляем набор первой вершиной графа.
4) Снова повторяем шаг 1.
5) Пусть новый конец имеет номер .
6) Обьявляем набор второй вершиной графа.
7) Продолжаем процедуру до прохода по всем точкам.
8) Ребра порождаем дугами.

Теперь считаем LCD случайной, т.е. полагаем вероятность каждой диаграммы равной где общее число диаграмм из (7). Т.о. мы получаем случайные графы. В [50] показано, что такие графы по своим вероятностным характеристикам почти неотличимы от графов . Графы с $n$ вершинами и $kn$ ребрами получаются так же, как и в модели Боллобаша-Риордана.

### 3.4. МодельBuckley-Osthus

Buckley, Osthus и другие исследователи предложили модификацию модели Барабаши-Альберт, в которой вершины обладают«изначальной привлекательностью» («initial attractiveness»): вероятность того, что старая вершина будет выбрана соседом новой вершины пропорциональна ее входящей степени(in-degree) плюс константе, т.е. «изначальной привле-кательности», т.е. $am$, где $m$ – число ребер, входящих в новую вершину. Если а , то мы просто получаем модель Бараба-ши-Альберт, т.к. там используется полная степень, а каждая исходящая степень равна m. Buckley иOsthus в работе [51] обобщили результат для произвольной привлекательности $a$.

Для произвольного фиксированного положительного целого $a$ граф определяется так же, как и в модели

Боллобаша-Риордана, но при этом вероятность выбора старой вершины определяется по-другому. Оказывается, в этой модели распределение степеней тоже подчиняется степенному закону. Если обозначить через число вершин графа с входящей степенью d, то при определенных условиях:

(3.11)

## 3.5. Модель копирования

### 3.5.1. Гененерация графа

Фиксируем и . В качестве начального графа возьмем любой *d*-регулярный граф. Пусть построен граф с номером *t*. Обозначим его через , где , причем *s* отличается от *t* на число вершин начального графа, т.е. на константу, выражаемую через *d*. Добавим к новую вершину и d ребер, выходящих из Для этого сначала случайно выберем вершину . Затем строим ребра из в На шаге с номером бросаем неоднородную монетку(падает решкой с вероятностью , орлом – с вероятностью ). Если выпала решка, то выпускаем ребро из в случайную вершину Если выпал орел, то берем *i*-го по номеру соседа *p*. Последнее действие всегда возможно, т.к. исходный граф *d*-регулярен.

### 3.5.2. Основной результат

**Теорема.** Пусть – математическое ожидание числа вершин степени *r* в графе . Тогда

. (3.12)

Также в модели копирования есть плотные двудольные графы, которые соответствуют спамерским структурам, отсутствующим в модели Боллобаша-Риордана.

## 3.6. Ориентированные безмасштабные графы

Большинство безмасштабных моделей описывают только неориентированные графы. Но поскольку большинство реальных сетей ориентировано, то логично создать ориентированные модели, в которых предпочтительное присоединение зависит от входящих и исходящих степеней. Такая модель была предложена Боллобашем, Риорданом и др [52]. В этой модели тоже получается степенной закон для входящих и исходящих степеней вершин.

## 3.7. Модель Чунг-Лу

### 3.7.1. Генерация графа

Пусть нам задано некоторое конечное множество вершин и степень каждой вершины . Генерация графа происходит следующим образом:

• Формируем множество $L$, состоящее из копий для каждого $i$ от 1 до $n$.

• Задаем случайные паросочетания на множестве $L$.

• Для вершин $u$ и $v$ из $Vs$ количество ребер в графе $G$, соединяющее их, равно числу паросочетаний между копиями $u$ и $v$ в $L$.

Сгенерированный таким образом граф соответствует степенной модели описывающей графы, для которых:

(3.13)

### 3.7.2. Основные результаты

• «Почти наверное» при граф связен.

• При в графе есть гигантская компонента, при этом все остальные компоненты имеют размер O (1).

• При в графе есть гигантская компонента, при этом все остальные компоненты имеют размер – решение уравнения .

• При меньшие компоненты имеют

• При в графе «почти наверное» нет гигантской компоненты.

### 3.7.3. Сравнение с реальными сетями

Авторы проверяли гипотезу на графе звонков (максимальное число узлов [3].

## 3.8. Модель Янсона-Лучака

### 3.8.1. Генерация графа

Рассмотрим упорядоченный набор из *ns* вершин. Каждой вершине *I* присваивается вес . Для простоты и понятности положим их независимо и одинаково распределенными случайными величинами со степенным хвостом:

(3.14)

с некоторыми константами и и некоторым Мы обозначаем наибольший вес Из (3.14) следует:

(3.15)

Следует отметить, что только если . В частности, при в случае распределения с экспоненциально неограниченным хвостом(heavy tail case) имеем .

При условии, что нам дан набор весов соединим кажду пару вершин посредством параллельных ребер, где – независимые пуассоновские случайные величины с ожиданием:

(3.16)

– константа. В результате мы получим мультиграф Далее мы можем стянуть параллельные ребра в одно, т.о. получив простой случайный граф , в котором вершины *i* и *j* соединены с вероятностью

(3.17)

независимой для всех пар таких, что

### 3.8.2. Основные результаты

Обозначим за размер максимальной клики в графе . В этих обозначениях в модели Янсона-Лучака получены следующие результаты.

**Теорема 5**
Если то

,

где

(3.18)

Если то

Если то почти наверное

Более того, при

(3.19)

(3.20)

Основной задачей при работе с веб-графом является поиск клик в нем.

Опишем несколько разных типов клик в модели Янсона-Лучака:

• Жадная клика (greedy clique)

• Quasi top clique

• Full top clique

Имеем . Обозначим за максимальную клику в графе.

Тогда

(3.21)

Тогда верны следующие теоремы.

**Теорема 6** Если то и имеют размер

С другой стороны,

(3.22)

**Теорема 7** Для любого существует алгоритм, который почти наверное находит в клику размером за полиномиальное время.

### 3.8.3. Основные результаты для схожих моделей

Вместо того, чтобы выбирать веса независимо согласно распределению $W$, можно выбрать их из подходящей детерминированной последовательности (как в модели Чунг-Лу), например

$$\bar{} \qquad ,i,\ldots,n. \qquad (3.23)$$

В данной модели верными остаются все результаты для модели Янсона-Лучака.

# 4. РАСПРОСТРАНЕНИЕ ЭПИДЕМИЙ В СОЦИАЛЬНЫХ ИНФОРМАЦИОННЫХ СЕТЯХ

## 4.1. Модель эпидемии SI

Сайты СИС в большинстве случаев не позволяют пользователям публиковать потенциально опасные файлы. Так большинство СИС дают возможность пользователям обмениваться видеороликами и фотографиями. Такие форматы данных не позволят встраивать исполняемый код и потому не могут содержать вирусы [54].

Но в СИС существует возможность обмена текстовыми сообщениями, которые могут содержать ссылки на вредоносные ресурсы. Факт повышенной доверчивости пользователей СИС позволяет использовать ее для распространения вирусов. Причем последние статистические сведения показывают высокий рейтинг успешности заражения таких вирусов [55, 56].

Следовательно, можно сделать вывод о существовании угрозы распространения вирусов в СИС. Данная угроза исходит от внутреннего, санкционированного пользователя, реализуется, используя существующие уязвимости СИС и создавая новый, с целью распространения вредоносного программного кода, способной нанести моральный вред.

Алгоритм реализации данной угрозы изображен на рис. 4.1.


Рис. 4.1. Угроза распространения вирусов в СИС

Распространение компьютерных вирусов в СИС во многом напоминает распространение инфекционных заболеваний в обществе. Таким образом, можно предположить, что модели описывающие процесс распространения эпидемий в обществе будут пригодны для описания процессов в СИС.

Наиболее простой моделью является модель SI (рис. 4.2, 4.3). В этой модели все вершины сети разделены на восприимчивые и инфицированные . В модели предполагается, что популяция имеет огромный размер (число вершин графа стремится к бесконечности). Инфицированные индивиды не выздоравливают и не умирают, а остаются зараженными [55,56].


Рис. 4.2. Переход состояния в SI модели

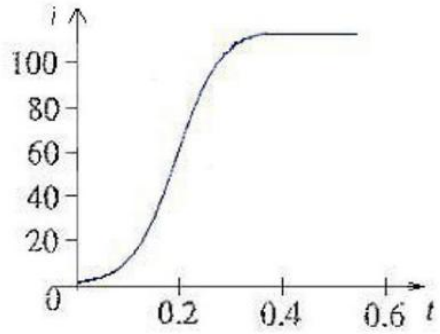



Рис. 4.3. График, соответствующий SI модели

Простейшая модель эпидемии для конечного числа элементов может быть представлена в виде выражения

—— ,

где ; – размер популяции (количество вершин графа); – коэффициент успешности заражения, или другими словами вероятность частного случая заражения.

Из выражения следует, что изменение числа инфицированных пропорционально коэффициенту успешности заражения , числу зараженных в момент времени *t* и числу оставшихся восприимчивых вершин .

На рис. 4.4 изображено распределение инфицированных и восприимчивых вершин. Модель SI не учитывает процесс исключения инфицированных вершин. Но для компьютерных вирусов это не приемлемо, так как необходимо учитывать возможность обновления системы и появления защиты от конкретной вредоносной программы. Но как бы то ни было, модель SI дает удобное представление о распространении заражения в начальные моменты времени, когда не существует «иммунитета» [54].

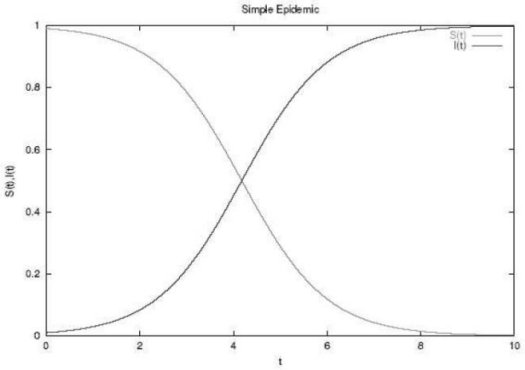



Рис. 4.4. Распределение зараженных и восприимчивых элементов сети

## 4.2. Модели просачивания и заражения

Модели просачивания (percolation) и заражения (contagion), представляют собой популярный способ изучения распространения информации. Классическая модель распространения эпидемии основана на следующем цикле заболевания носителя: первоначально человек восприимчив к заболеванию (susceptible); если он входит в контакт с инфицированным, то заражается (infected & infectious) с некоторой вероятностью ; впоследствии через некоторый период времени человек становится здоровым, приобретая иммунитет, или умирает (recovered/removed); иммунитет со временем снижается, и человек снова становится восприимчивым к болезни (susceptible) [54].

В модели SIR (по первым буквам трех этапов цикла заболевания) [55, 56] выздоровевший становится невосприимчивым к болезни: .

Соответственно общество представляется тремя группами (рис. 4.5, 4.6):

— численность группы людей, еще не инфицированных или восприимчивых к болезни в момент времени ;

— численность группы инфицированных людей;
— численность группы выздоровевших людей.

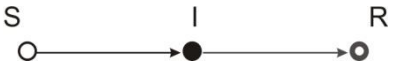



Рис. 4.5. Переход состояния в SIR модели

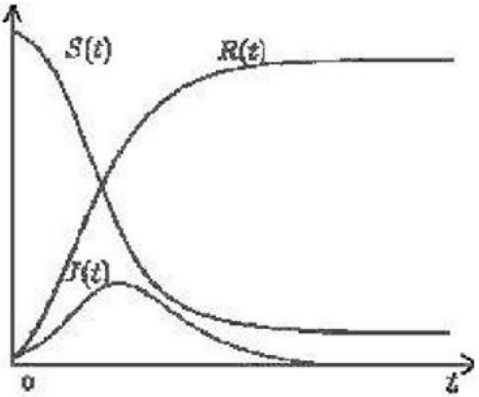



Рис. 4.6. График, соответствующий SIR-модели

Пусть
Динамика следующая:

т.е. каждый из инфицированных в единицу времени, контактируя с восприимчивыми к болезни, заражает их с вероятностью .

.

Инфицированные выздоравливают через средний период времени , соответственно,

Существуют и другие аналогичные более сложные модели, в частности, в модели SIRS выздоровевший становится восприимчивым к болезни через некоторое время.

Простейший пример ситуации, где такая модель является естественной – распространение информации в социальной сети. Блоггер (человек, который ведет блог – сетевой дневник) может прочитать блог друга (восприимчив), посвященный некоторой теме, а затем может и сам написать об этой теме (инфицирован), и позже вернуться к ней (восприимчив) [55,56].

Для социальных сетей ключевым показателем является «эпидемический порог» – критическая вероятность заражения соседа, при превышении которой «инфекция» распространяется по всей сети. Эпидемический порог зависит от свойств графа социальной сети, например: числа вершин, распределения связей, коэффициента кластеризации. Поэтому распространение инфекции сильно зависит от выбранной модели представления графа сети.

Если социальную сеть представить случайным графом, то инфекция с вероятностью заражения выше порога экспоненциально быстро размножается

Инфекция, с вероятностью заражения ниже порога, экспоненциально быстро «вымирает».

Более реалистичной моделью социальной сети является безмасштабный граф, в котором некоторые вершины связаны с тысячами и даже миллионами других вершин, в большинстве своем имеющих всего по нескольку связей (т. е. отсутствует

характерный масштаб). В таком графе распределение количества связей узлов описывается степенным законом [57].

Анализ распространения компьютерных вирусов в безмасштабных сетях показал, что в них эпидемический порог отсутствует — эпидемия охватит всю сеть, если возникнет инфекция [58]. Однако в блогосфере многие обсуждаемые темы могут распространяться без возникновения эпидемий, поэтому порог все же отличен от нуля [59,60].

## 4.3. Модель распространения эпидемии, адаптированная к социальным информационным сетям

Для создания правдоподобной модели нам необходимо учитывать процесс распространения вируса в сети. Распространение заражения начинается с момента, когда злоумышленник создал некоторый вредоносный код и разместил в сети, в этом случае мы понимаем первичное заражение нескольких элементов сети. Вирус начинает свободно распространяться, заражать другие элементы сети. Через некоторое время вирусная активность обнаруживается производителями антивирусных программ. С этого момента начинается работа, направленная на изоляцию вируса и разработку сигнатуры вируса. Сигнатура будет использоваться антивирусными программами для дальнейшего обнаружения вируса. Процесс создания сигнатуры также занимает некоторое время, в течение которого вирус продолжает свободно распространяться [61].

Следующим этапом в борьбе с вирусом является распространение «вакцины». После того, как была разработана сигнатура для обнаружения, ее необходимо отправить всем элементам (компьютерам) сети, необходимо разослать ее миллионам, а в случае с СИС и сотням миллионов [55, 56].

После того, как некоторый элемент сети (компьютер) получил сигнатуру, возникают два возможных состояния: к этому моменту элемент сети еще не был инфицирован, и тогда он приобретает иммунитет в виде имеющейся сигнатуры, или

же к моменту получения сигнатуры он уже был инфицирован, тогда пользователь может обнаружить и обезвредить вирус. Обычно рекомендуется отключить компьютер от сети Интернет, если был обнаружен вирус, что способствует предотвращению дальнейшего распространения вируса. После того, как вирус был обезврежен, компьютер можно снова подключить к сети Интернет [54].

Таким образом, переходы состояния каждой отдельной вершины графа можно изобразить в следующем виде (рис. 4.7).

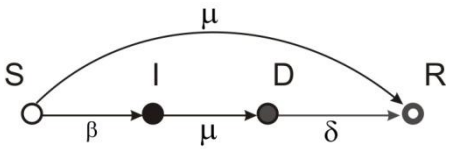



Рис. 4.7. Переход состояния элементов модели распространения вируса в СИС

В таком случае мы предполагаем, что каждая вершина может находиться в одном из четырех состояний: восприимчивая , инфицированная , обнаружено заражение и восстановлена или удалена Эта модель называется Progressive Susceptible Infectious Detected Removed . До момента повеления сигнатуры вирус может свободно распространяться, и тогда будут присутствовать только два возможных состояния и (рис. 4.8).

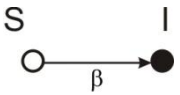



Рис. 4.8. Процесс заражения

Таким образом, в нашей модели должны быть объединены как SI модель, так и SIDR модель.

Как говорилось ранее, при моделировании процесса распространения мы используем показатель , который

отображает число успешный заражений за один такт времени. Мы предполагаем, что всего пройдет таких тактов до момента обнаружения вируса и создания сигнатуры. Таким образом, можно сказать, что при мы имеем модель SI, а при - модель SIDR.

Для описания процесса распространения сигнатуры мы используем параметр , который отображает число компьютеров, которые получат «вакцину» за один такт времени. Процесс «вакцинации» независим от состояния элемента, и если в момент получения «вакцины» элемент находился в состоянии , то он перейдет в состояние , а если он находился в состоянии , то перейдет в состояние .

Процесс восстановления (удаления) вершин, у которых было обнаружено заражение (состояние ), описывается с помощью параметра .

В нашей модели используется дискретное представление времени, оно разделяется на ряд отдельных тактов, и в каждом таком такте все множество элементов сети в определенном состоянии изменяется в соответствии с определенным правилом [54].

Рассмотрим модель подробнее. Предположим, что в начальном состоянии только один элемент сети является зараженным , а все остальные – восприимчивые . До момента обнаружения и создания сигнатуры проходит время равное . В это время вирус распространяется и инфицирует восприимчивые элементы. Вероятность того, что некоторая восприимчивая точка графа станет зараженной, определяется в соответствии с выражением

где – коэффициент успешности заражения; – число инфицированных элементов; – общее число всех элементов сети.

В данном случае параметр позволяет приблизиться к существующей статистике и создать real-world модель [55, 56].

После того, как вирус обнаруживается и создается «вакцина» , начинает распространяться сигнатура вируса и зараженные элементы восстанавливаются.

Модель приобретает новый вид:

Основное преимущество такой модели заключается в том, что она позволяет оценить размер эпидемии, и соответственно ее последствия. Spam-сообщения могут значительно повысить трафик в сети. Это в свою очередь приведет к снижению производительности сервиса или вообще к отказу в обслуживании. Также это затронет репутацию конкретной СИС, что является недопустимым ущербом для большинства из них [61].

# 5. ИНФОРМАЦИОННЫЕ РИСКИ И ЭПИСТОЙКОСТЬ БЕЗМАСШТАБНЫХ СЕТЕЙ

Большие и сверхбольшие [62] сетевые структуры стали весьма популярным ресурсом современного информационного общества. Их востребованность порождает значительные информационные риски, анализ которых сегодня является неотъемлемым атрибутом процесса обеспечения безопасности сетей [2-5]. Социальные сети [62] как многие другие сетевые структуры имеют безмасштабную архитектуру [67-70]. В этой связи представляется актуальным исследование безмасштабных сетей на предмет [71, 72] распространения в них вредоносного программного обеспечения (ВПО).

Будем исходить из того, что безмасштабная сеть (рис. 5.1) атакована ВПО (например, компьютерным вирусом) в её *k*-слое, т.е. атакован хотя бы один элемент сети степени *k*. Согласно свойству безмасштабной сети доля (в общем множестве элементов сети) узлов $k$ – степени (имеющих $k$ коммуникаций в нутрии неё) составляет , где $\gamma = 2 \div 3$.Пирамидальная топология безмасштабной сети (чем больше у узла инциденций, тем меньше узлов этого качества) позволяет предположить, что эпидемия будет, скорее всего, развиваться по направлению к основанию пирамиды (рис. 5.1), т.е. от $k$ - слоя к – слою и т.д. Пошагово процесс заражения ВПО иллюстрирует рис. 5.2. Принципиально возможным здесь следует считать два параметра: общее количество узлов участвующих в эпидемии, и количество пораженных ВПО узлов на каждом шаге $i$ процесса.

Временной шаг процесса определяется инкубационным периодом ВПО, т.е. интервалом времени, в течении которого инфекция проявляет себя в слое негативными последствиями (сбоями и отказами в их работе). Следуя предлагаемой модели (рис. 5.2), можно определить основные параметры процесса на каждом его шаге. Так для общего количества узлов, участвующих в эпидемии, имеем пошагово:

шаг 1 :N[1]=1;

шаг 2 :N[2]=1+(k-1);
шаг 3 :N[3]=1+(k-1)+(k-1)(k-2);
…
шаг r :N[r]=1+ ———.

Для оценки ожидаемого множества пораженных узлов воспользуемся функцией , оценивающей ожидания заражения узлов на любом этапе *i* эпидемии согласно соответствующем законам распределения вероятностей. В зависимости от потребностей анализа это может быть: матожиданием (усредненные оценки), мода а (пиковые оценки), матожидание с дисперсией m±σ (диапазонные оценки) и т.п.

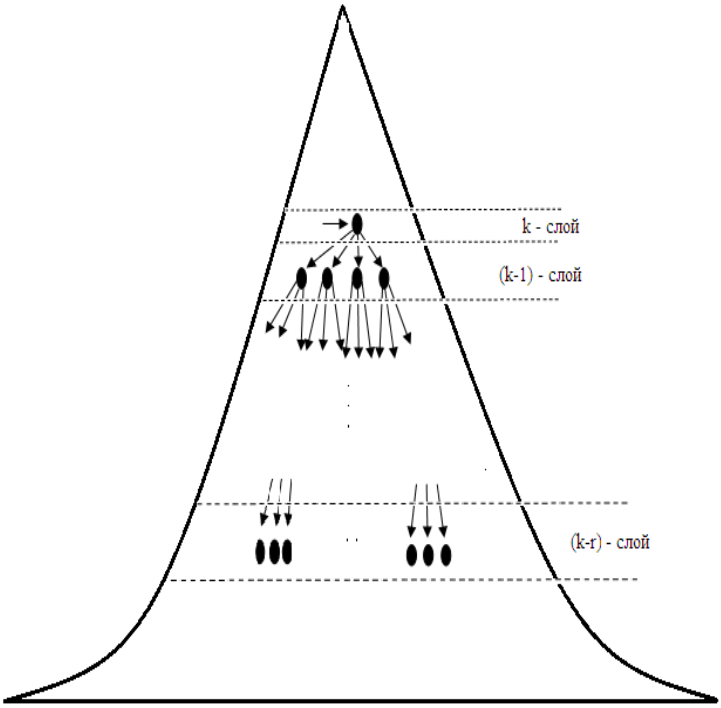



Рис. 5.1. Иллюстрация вирусной эпидемии
в безмасштабной сети с учетом узлов

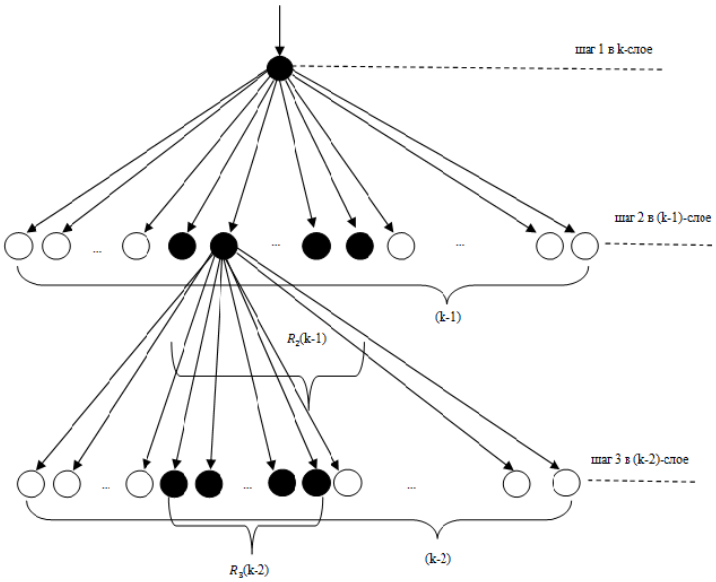



Рис. 5.2. Иллюстрация вирусной эпидемии в безмасштабной сети с учетом пошагового заражения

Для дискретных распределений (биномиального, гипергеометрического, Паскаля, Пуассона и др.) фактически это первичные меры риска, показывающие в целочисленном выражении долю зараженных узлов из общего числа атакуемых. Например, при атакуемое множество узлов имеет мощность , а ожидаемое количество пораженных узлов будет равно 2 (мода) или 3 (матожидание), при дисперсии 1 в диапазоне (1÷3).Для различных слоев модели эти оценки будут иметь свои аналитические выражения и параметры. Реализуя эту процедуру послойно, можно получить необходимую общую модель эпидемии в сети.

Что же касается параметра r, то он характеризуется длительностью свободного развития эпидемии, например, до момента ее обнаружения и принятия, соответствующих мер противодействия.

Распишем по шагам ожидаемое количество зараженных узлов (рис. 25):

шаг1 : I[1]=1;

шаг2 : I[2]=1+ ;

шаг3 : I[3]=1+ + ;

…

шаг r : I[r]= 1+

где R – уязвимость, т.е. параметр характеризует ожидаемое количество зараженных узлов в *j*-ом слое от вирусованного узла в предыдущем слое.

В случае, если ожидаемое количество зараженных узлов на каждом слое (шаге) процесса равно ,j=1(1)r, имеем:

$$I[r] = 1 + R + \ldots + = \text{——}.$$

В свою очередь общий риск на i-ом шаге будет равен отношению ожидаемого количества зараженных узлов к общему количеству узлов, подверженных воздействию ВПО, т.е.

$$Risk[i] = \text{—}. \qquad (5.1)$$

Эпистойскость сети можно оценить как отношение ожидаемого количества незараженных узлов и их общему количеству, т.е.

$$L[i] = \text{———} = 1 - Risk[i]. \qquad (5.2)$$

Далее остается найти все вышеуказанные аналитические выражения для различных распределений и мер риска.

Важным исходным моментом является нахождение нормирующего множителя C в безмасштабном распределении

$$P(k) = \frac{}{} = \frac{}{}, \qquad (5.3)$$

где - количество узлов сети степени k;

N – общее количество узлов сети.

При этом очевиден инвариант

$$\frac{}{} \quad \text{или} \quad C \quad \frac{}{} \quad .$$

Отсюда

$$C = \frac{}{}.$$

В отношении распространенного на практике случая, когда $\gamma=2$, уместно вспомнить сумму ряда

$$S = 1 + \frac{}{} \quad \frac{}{} \qquad \frac{}{} \qquad \frac{}{}.$$

При достаточно больших k□1000 интересно отметить, что эта сумма приближается к «золотому сечению».

Для $\gamma=2$ ненормированный параметр составит $C \approx 0.62$. Уместно в этой связи рассмотреть следующий пример, где k = 2(1)10 (табл. 5.1)

Таблица 5.1

Пример

| k | 10 | 9 | 8 | 7 | 6 |
|---|---|---|---|---|---|
| P(k)/C | 0.0100 | 0.123 | 0.156 | 0.0204 | 0.0277 |

| k | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| P(k)/C | 0.0400 | 0.0625 | 0.1111 | 0.2500 | 1.1000 |

Отсюда имеем:

.

Причем . Тогда получаем нормированный параметр, равным:

C =

С ростом k знаменатель, очевидно, будет нарастать, и предел будет приближаться к значениям «золотого сечения».

Открытым остался вопрос о пределах суммирования P(k). В этой связи следующее предельное уравнение:

или

,

где [] – оператор вычисления целой части.

Таков теоретический максимум степени вершин безмасштабной сети, включающей N узлов. Так для γ=2, N=162 и C=0.62 получаем для вышерассмотренного примера.

### 5.1. Риск-факторы безмасштабной сети

Рассмотрим процесс распространения вредоносного программного обеспечения (ВПО) в безмасштабной сети. Целенаправленная атака обычно предполагает деструктивное воздействие на вершины с максимальными степенями. Общее количество вершин с максимальными степенями может быть получено из закона распределения степеней вершин в безмасштабной сети (5.3).

Вероятность успеха единичной атаки на узел степени k может быть оценена по-разному. Так можно оценить вероятность успешной атаки на один узел каждой степени в сети и данное значение распространить на все узлы с заданным значением степени вершины. Исходя из этого, для безмасштабного распределения можно рассчитать уязвимость, т.е. ожидаемое количество зараженных вершин степени по формуле:

(5.4)

где – количество вершин со степенью k;

– вероятность заражения вершины со степенью k.

Используя выражение (4), можно оценить риск одновременной атаки на узлы степени k и возможность отказа из них. При этом рассматривается только эффект, предполагающий, что атака не распространяется на смежные с атакуемыми узлы (ветвящийся процесс). Для устранения данного недостатка предлагается использовать распределение вершин по слоям (рис. 5.3), подразумевающее, что атакованные узлы k-ого слоя являются источниками атаки на слой (k-1).

После проведения целенаправленной атаки на s вершин со степенью k можно провести атаку на вершины, связанные с s вершинами, лежащими на слое (k-1). Для реальной сети можно определить вершины связанные с данной, однако для теоретически заданной сети связь данной вершины с другими носит вероятностный характер, и, в большинстве случаев, нельзя определить вершину, связанную с данной. В связи с этим будем предполагать, что атакуемая вершина связана с вершиной с максимальной степенью, а все остальные ребра идут либо к вершинам со степенью, совпадающей со степенью атакованной вершины, либо с вершинами, имеющими меньшую степень (рис. 5.4).

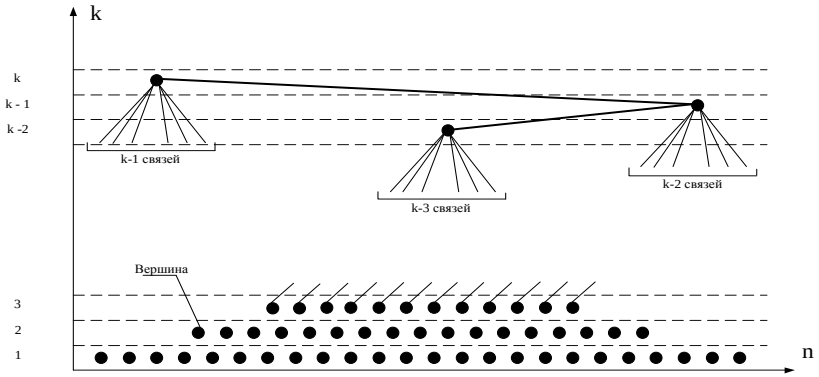



Рис. 5.3. Распределение вершин по слоям

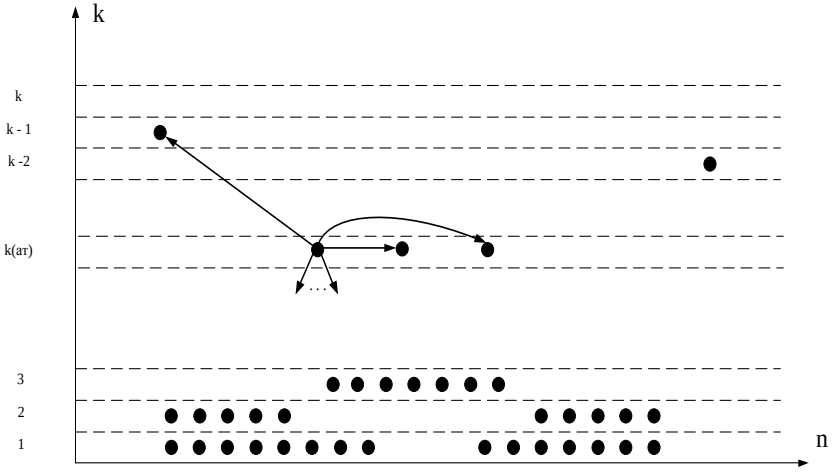



Рис. 5.4. Нецеленаправленная атака в безмасштабной сети

## 5.2. Определение ущерба

Для определения ущерба при атаке вершин k-слоя воспользуемся следующей моделью:

, (5.5)

где – средний ущерб при отказе атакуемых узлов на k-слое;

– длительность инкубационного периода и время простоя успешно атакованных ВПО узлов сети;

T – период рассмотрения и время простоя успешно атакованных ВПО узлов сети;

– средняя пропускная способность каналов связи для узла на k-слое;

– средняя ценность информации, циркулирующей в единицу времени на k-слое;

– уязвимость узлов в k-слое, атакуемых рассматриваемым ВПО (ожидаемое количество отказавших узлов).

Используя (5.5), можно найти ущерб, наносимый ВПО компьютерной сети, как сумма ущербов по всем слоям безмасштабной сети:

(6)

Параметр характеризует длительность свободного развития атаки ВПО, например до момента ее обнаружения и принятия соответствующих мер противодействия.

Для представленной модели ущерба проведем оценку пропускной способности каналов в k-слое безмасштабной сети.

Чтобы провести данный анализ введем некоторые ограничения на систему передачи информации (СПИ) [71], а именно: передача информации происходит без предварительного согласования между узлами сети и без подтверждения о приеме информации. При данных ограничениях сеть представляет собой одноканальную систему

массового обслуживания с отказами. Тогда СПИ может находиться в одном из двух состояний:

- S0 – канал свободен для передачи информации;
- S1 – в канале передается информация.

Эффективность при такой работе зависит от интенсивности λ поступления заявок на обслуживание (передачу информации) и интенсивность μ обслуживания заявок (передача информации). Полагая, что данные потоки являются простейшими, в стационарном режиме система уравнений Колмогорова для вероятностей состояний вырождается в уравнение (5.7):

(5.7)

Исходя из того, что сумма вероятностей событий должна быть равна 1, т.е.

,

вероятность готовности к обслуживанию может быть определена как:

— (5.8)

а вероятность отказа:

— (5.9)

Для простейшего потока запросов вероятность поступления m заявок ( ) за время t можно найти по закону Пуассона:

—— (5.10)

В данном случае вероятности распределены с параметром .

Проведем оценку интенсивности потока обслуживания. Пусть интенсивность потока обслуживания в канале без ошибок равна:

——— (5.11)

где – среднее число запросов, адресованных к одному узлу k-слоя;

– средний объем информации, передаваемой на один узел k-слоя при одном запросе;

– средняя скорость обработки информации, на одном узле k-слоя;

– число элементов на k-слое.

Зададим интенсивность входного потока на k-слой. Несложно заметить, что интенсивность входного потока должна зависеть от объема информации, поступающей на k-слой за определенный период, отсюда можно задать интенсивность входного потока как:

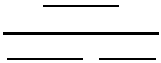
——, (5.12)

где - объем информации, поступающий на k-слой за период .

Учтем вероятность правильного приема ( ) пакета данных в канале связи с вероятностью ошибочного приема информации ( ). Тогда . Относительная пропускная способность СПИ на k-слое с учетом передачи информации с ошибками равна:

(5.13)

Преобразуем формулу (5.13), подставив значения (5.11) и (5.12). Получим следующую формулу (5.14):



(5.14)

Формула (5.14) позволяет определить среднюю пропускную способность для каналов связи узлов на k-слое в зависимости от числа запросов, поступающих на k-слой, объема информации, средней скорости обработки и вероятности ошибочного приема информации. Это возможно для оценки ущерба с точки зрения количества циркулирующей в сети информации.

Вместе с тем для определения размера ущерба необходимо учитывать и ценность утраченной в результате атаки информации.

Существует несколько подходов к оценке ценности информации. Наибольшее распространение получили затратный и вероятностный способы определения ценности информации.

Принцип оценки ценности информации основывается на следующем. Если известно, что цель наверняка может быть достигнута и притом несколькими путями, то возможно определение ценности информации S, например, по уменьшению материальных или временных затрат, благодаря использованию информации. В общем виде это можно выразить формулой:

, (5.15)

где S – количественная мера ценности информации, например, в денежном измерении ($) или временном (Т);

n – возможное число путей решений задачи по минимизации материальных затрат;

m - возможное число путей решений задачи по минимизации временных затрат.

Применение такого метода оценки ценности информации для безмасштабных сетей является затруднительным, т.к. определение параметров n и m требует оценки, исходя из конкретики информации, поступающей на k-слой.

Для оценки ценности информации при помощи вероятностного способа будем считать, что P — вероятность достижения цели после получения информация, p — вероятность достижения цели до получения информации. Используя данные понятия можно предложить следующую оценку меры ценности:

— (5.16)

Применение вероятностного способа оценки ценности информации для безмасштабных сетей является проблемным, т.к. требует определения вероятностей достижения цели с учетом поступившей информации и без нее, что является крайне трудоемким для всего массива поступающей информации.

Предложенные выше подходы либо не могут быть использованы для оценки ценности информации в слоях безмасштабной сети, либо полностью основываются на субъективных взглядах эксперта. В связи с этим предложена следующая модель оценки ценности информации.

Можно исходить из того, что чем больше степень узла безмасштабной сети, тем более ценная информация обрабатывается в узле (устранение узла приведет к ликвидации наибольшего числа ребер в сети; чаще всего узлы с наибольшими степенями являются узлами управления сетью; устранение узла приведет к нарушению топологии сети, и, как следствие, к перераспределению маршрутов информации, что приведет к увеличению затрат и увеличению времени

прохождения информации). Можно сказать, что ценность информации в безмасштабной сети зависит так же от времени актуальности информации. Аналогично с моделями определения класса информации (ГРИФ) в безмасштабных сетях можно ввести понятие класса информации для слоя или нескольких слоев, определяющие относительную ценность информации, циркулирующей на слоях. Исходя из всего вышеперечисленного, можно вывести эмпирическую формулу ценности информации для k-слоя:

(5.17)

где: – степень вершин слоя;

– время актуальности информации, поступающей на k-слой;

– класс информации.

Будет считать, что ценность информации является относительной величиной, изменяющейся в пределах от 0 до 1. Исходя из формулы (5.17) можем определить ценность информации по формуле (5.18):

(5.18)

где: П – приоритет обработки запроса, поступающего на k-слой;

h – количество запросов, поступающих на k-слой.

Использование формулы (5.18) позволяет просуммировать относительные ценности каждого запроса, поступающего на k-слой, а также учитывает приоритет обработки запроса.

Введение приоритета позволяет строить динамическую модель обработки запросов, учитывающую передачу информации с одного слоя на другой, ценность и срочность обработки информации, определяемые на каждом шаге её обработки. Исходя из вышесказанного, определим приоритет для i-ого запроса:

——— (5.19)

где: – базовый приоритет i-ого запроса, определяющийся уровнем k-слоя на котором он был создан;

– коэффициент добавочного приоритета, определяемый пользователем, обрабатывающим данный запрос (лежит в диапазоне от —— до ——);

– коэффициент актуальности информации, определяемый пользователем, обрабатывающим данный запрос (лежит в диапазоне от 0 до 1);

– время ожидания обработки запроса;

–время обработки запроса.

Проведем анализ предложенной модели оценки ценности и задания приоритета запроса.

Рассмотрим формулу (5.18). При любых значения П и g значение не превышает 1. Значение 1 будет достигнуто в том случае, когда информация, поступающая на k-слой, будет иметь максимальный класс и максимальный приоритет. Данную формулу можно расширить, добавив в формулу параметр, который дает финансовую оценку стоимости информации.

Проведем анализ формулы (5.19). Базовый приоритет задается всем запросам, создаваемым на k-слое. Данный коэффициент может быть задан несколькими способами. При максимальном упрощении подхода, можно значение приоритета приравнивать к значению k. В этом случае количество базовых приоритетов будет совпадать с количеством слоев. С целью увеличения точности оценки базовый приоритет на k-слое можно задавать в диапазоне ( ), где Δ – параметр, не превышающий базовой оценки приоритета. Это позволяет при создании запроса предоставить возможность передавать информацию либо на слой, стоящий выше или ниже для слоя, создающего запрос. Однако, это может увеличить разброс приоритета, что может

привести к передаче запроса не на ограниченное количество слоев от слоя создания (заранее заданный диапазон), а на некоторое, значительно большее значение. Это может привести к нарушению «субординации» запросов. В связи с этим, будем считать, что при создании запроса базовый приоритет не меняется. С целью задания ограниченного разброса приоритета (для передачи на слои выше или ниже) введем коэффициенты и .

Коэффициент задается создателем запроса. Так как диапазон коэффициента зависит от базового приоритета (от —— ——), то базовый приоритет за счет данного коэффициента может быть увеличен или уменьшен максимум на 1. Это позволяет регулировать общий приоритет запроса.

С целью введения зависимости приоритета от времени актуальности информации введен коэффициент . Введение коэффициента позволяет более гибко подходить к оценке приоритета в зависимости от времени актуальности запроса. Это позволит при минимальном времени актуальности увеличивать приоритет, если адресат запроса лежит не на слой выше данного, а на некоторую другую величину.

Оценим разброс приоритета в зависимости от коэффициентов, задаваемых создателем запроса. При значениях коэффициента равных —— —— приоритет может быть увеличен или уменьшен на 1. При любых значениях коэффициента значение ———— лежит в диапазоне от 0 от 1.

Исходя из предыдущих оценок, можно сказать, что диапазон приоритета при создании запрос лежит в границах .

Для каждого уровня k должен быть задан класс информации . Введение данного значения позволяет заранее определить диапазон, в котором будет изменяться ценность передаваемой информации (аналогом класса информации может быть разделение информации по грифу секретности.

Класс информации можно задать как число, исходя из которого, при любом значении приоритета обработки запроса значение ценности информации не опускалось бы ниже обозначенного класса информации.

Предложенные модели ценности и задания приоритетов позволяют максимально гибко подойти к вопросу оценки ущерба. Вышеуказанная зависимость от класса информации, времени обработки и слоя безмасштабной сети позволяет провести максимально эффективное ранжирование информации при знании базовых данных о ней. Также предложенные модели позволяют учесть особенности безмаштабных сетей.

## 5.3. Оценка мощности ожидаемого множества успешно атакованных узлов

Проведем оценку ожидаемого множества пораженных ВПО узлов с использованием функции Ri, оценивающую количество успешно атакуемых узлов на шаге i согласно закону распределения вероятностей.

Рассмотрим дискретные распределения (биномиальное, отрицательное биномиальное, Пуассона, Бернулли, геометрическое, гипергеометрическое, логарифмические) в соответствии с которыми будем определять долю успешно атакованных узлов[69,70].

Для заданных распределений имеем следующие параметры (табл. 5.2):

Таблица 5.2

Характеристики дискретных распределений случайных величин

| Распределение | Функция вероятности | Мат. ожидание | Мода | Дисперсия |
|---|---|---|---|---|
| Биномиальное | | np | [np] | npq |
| Пуассона | | $\lambda$ | [$\lambda$] | $\lambda$ |
| Бернулли | | p | 0 или 1 | pq |
| Отрицательное биномиальное | | — | ——— | — |
| Геометрическое | | – | 0 | — |
| Логарифмическое | ——— | ——— | 1 | ——— |

Выведем значение I[r] и для каждого распределения, подставляя которое в формулы (5.1) и (5.2) можно оценить риск и эпистойкость.

Используя формулу (5.4) подставляя значения математического ожидания биномиального распределения вместо произведения степеней вершин на в I[i], можно получить ожидаемое количество успешно атакованных узлов на каждом шаге:

шаг1 : I[1] = 1;
шаг2 : I[2] = 1 + ;
шаг3 : I[3] = 1 + + ;
(5.20)
…
шаг r (для ) : I[r] =1+ .

Исходя из (5.20) можно получить оценку для риска и эпистойкости:

(5.21)

(5.22)

Формулы (5.21) и (5.22) позволяют получить оценку для риска и эпистойкости при биномиальном распределении качества успешно атакованных вершин на произвольном шаге процесса распространения ВПО в сети.

Для проведения оценки риска и эпистойкости при отрицательном биномиальном распределении необходимо ввести параметр hk, определяющий количество успешно атакованных узлов степени k из общего количества узлов на шаге i. Согласно табл. 5.2 мат. ожидание данного распределения равно:

—.

Подставим в данную формулу значение $h_k$ вместо r, вероятность появления узла степени k определим как .В результате имеем:

шаг1 : I[1] = 1;

шаг2 : I[2] = 1 + ————————;

шаг 3: I[3] = 1 + ———— +————————————; (5.23)

…

шаг r (для ); I[r]=1+ + ————————————

Исходя из формулы (23) можно получить оценку для риска и эпистойкости:

————————————, (5.24)

———————————— (5.25)

Выражения (5.24) и (5.25) дают оценку риска и эпистойкости при отрицательном биномиальном распределении количества успешно атакованных вершин.

Определим, что количество выведенных ВПО из строя узлов на каждом шаге ( ) определяется по формуле Пуассона с заданным значением λ:

. (5.26)

Математическое ожидание распределения Пуассона равно λ, таким образом будем считать, что на каждом шаге будет выведено из строя λ узлов:

шаг 1: I[1] = 1;
шаг2 : I[2] = 1 + λ* ;
шаг3 : I[3] = 1 + λ* + λ* ;
…
шагr : I[r] =1+ (5.27)

Исходя из формулы (5.27) можно получить следующую оценку для риска и эпистойкости:

(5.28)

(5.29)

Выражения (5.28) и (5.29) позволяют получить аналитическую оценку риска и эпистойкости при отрицательном пуассоновском распределении количества успешно атакованных вершин на произвольном шаге r развития эпидемии.

Далее для оценки ожидаемого количества выведенных из строя количества узлов степени k используем математическое ожидание логарифмического распределения. По аналогии с вышеизложенным имеем:

шаг1 : I[1] = 1;
шаг2 : I[2] = 1 + ; (5.30)
шаг3 : I[3] = 1 + + ;
…

шагr (для ): I[r] = 1+ .

Отсюда можно получить аналитическое выражение для риска и эпистойкости:

(5.31)

. (5.32)

Выражения (5.31) и (5.32) позволяют осуществить оценку риска и эпистойкости при логарифмическом распределении успешно атакованных вершин.

При геометрическом распределении вероятностей вывода вершин из строя имеем среднее число успешно атакованных вершин на каждом шаге:

шаг1 : I[1] = 1;

шаг2 : I[2] = 1 + ; (5.33)

шаг3 : I[3] = 1 + + ;

…

шаг r (для ): I[r]=1+ .

Из (5.33) имеем аналитическое выражение для риска и эпистойкости при геометрическом распределении количества успешно атакованных ВПО вершин:

(5.34)

(5.35)

Выражения (5.34) и (5.35) позволяют сделать численную оценку для риска и эпистойкости при геометрическом распределении количества успешно атакованных ВПО вершин на любом шаге развития эпидемии. Вышеприведенные риск-оценки базируются на матожидании дискретных распределений, но они могут быть расширены для случаев пиковых (мода) и диапазонных (матожидание и дисперсия) оценок.

## 5.4. Подходы к управлению эпистойкостью атакуемой безмасштабной сети

В отношении безмасштабных сетей оценку живучести эффективно провести на основе оценки эпистойкости сети. Исследуем формулу нахождения эпистойкости (5.2):

—

Знаменатель дроби является постоянной величиной для i-ого шага. Соответственно для увеличения эпистойкости необходимо, чтобы числитель – количество незараженных узлов, принимал максимальное значение. Рассмотри формулу нахождения :

——— (5.36)

В формуле (5.36) единственным элементом, не являющимся константой, является функция Ri, оценивающая количество успешно атакуемых узлов на шаге i. Значение будет принимать максимум при минимуме Ri.

Значение функции Ri определяется законом распределения вероятности. Однако, для любого из рассмотренных распределений, вводится зависимость от – вероятности заражения вершины со степенью k.

Исходя из вышесказанного необходимо, чтобы параметр принимал как можно меньшие значения.

Практическая реализация по уменьшению вероятности заражения узлов может сильно различаться в зависимости от принятой стратегии защиты. Использование стратегии максимальной защиты узлов с наибольшими степенями не является оптимальной, т.к. приводит к тому, что на защиту элементов меньшей степени не отводится достаточно средств. В связи с этим воздействие ВПО будет эффективно для большей части узлов сети, что приведет к уменьшению эпистойкости, а как следствие, нарушит функционирование распределенной компьютерной сети.

Использование стратегии равномерной защиты всех узлов сети так же не является оптимальной, т. к. в случае проведения целенаправленных атак (данный тип атак является наиболее распространённым) с большой вероятностью будут выведены из строя узлы с максимальными степенями, что приведет к устранению наибольшего числа ребер в сети. Следствием этого станет нарушение её топологии. Для ликвидации последствий данной атаки необходимо будет провести перераспределение маршрутов информации, что приведет к увеличению затрат и времени прохождения информации. Также, вывод из строя узлов с наибольшими степенями является максимально неприятным с точки зрения финансовых затрат, полученных из-за простоя сети (нарушения корректности функционирования) и средств, которые необходимо потратить на восстановление работоспособности данных узлов.

Квазиоптимальным представляется использование стратегии обеспечения приемлемого уровня риска для узлов с маленькими степенями и высокого уровня живучести для узлов с большой степенью, однако данная стратегия требует значительных ресурсов.

Оптимальная стратегия, обусловленная топологией и распределением ресурсов на защиту сетевых элементов, состоит в оценке стратегии и действий атакующей стороны.

Однако применение данной стратегии приемлемо для типовых атак. Вредоносные воздействия, основанные на новом принципе (появление новых угроз является обыденным для современного уровня технического развития) приведет к краху сети.

# ЗАКЛЮЧЕНИЕ

В данном учебном пособии автор поставил своей целью осветить теоретические аспекты и развить научно-методические подходы к обеспечению информационной безопасности социальных сетей в плане исследования атак на информационные технологии и системы в следующих основных аспектах:

- изложение основных топологических особенностей безмасштабных сетей в контексте проведения в них информационных атак
- изложение специфики оценки живучести безмасштабных сетей на различных стадиях развития информационной атаки, включая моделирование процесса и выработку рекомендаций по управлению живучестью РКС
- описание методологии моделирования сетевых структур
- развитие методологии анализа математических моделей заражения вредоносным программным обеспечением компьютеров пользователей социальных информационных сетей. Были рассмотрены существующие модели эпидемий. На их основе была построена модель распространение вирусов в СИС. Полученная модель дала возможность оценить факторы, влияющие на размер ущерба наносимого вирусной эпидемией в СИС.

Настоящий материал имеет научную и практическую ценность для студентов старших курсов специальностей направления «Информационная безопасность», а также аспирантов научной специальности 05.13.19 «Методы и системы защиты информации, информационная безопасность».

## БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Чураков, А. Н. Анализ социальных сетей [Текст] / А. Н. Чураков // СоцИс. – 2001. – № 1. – С. 109–121.

2. Charu, C. Aggarwal [Text] / C. Charu. – Social Network Data Analytics, 2011. – 520 p.

3. Milgram, S. The Small World Problem [Text] / S. Milgram // Psychology Today. – 1967. – Vol. 2. – P. 60–67.

4. Granovetter, M. S. The Strength of Weak Ties [Text] / M. S. Granowetter // American Journal of Sociology. – 1973. – Vol. 78. – No. 6. – P. 1360–1380.

5. Kleinberg, J. M. Authoritative Sources in a Hyperlinked Environment [Text] / J. M. Kleinberg // J. ACM. – 1999. – Vol. 46. – No. 5. – P. 604–632.

6. Johnson, J. Assessing Children's Sociometric Status: Issues and the Application of Social Network Analysis [Text] / J. Johnson, M. Ironsmith // Journal of Group Psychotherapy, Psychodrama & Sociometry. – 1994. – Vol. 47. – Is. 1. – P. 36–49.

7. Gyöngyi, Z. Combating Web Spam with TrustRank [Text] / Z. Gyöngyi, H. Garcia-Molina, J. Pedersen // Proceedings of the International Conference on Very Large Data Bases. – 2004. – Vol. 30. – P. 576.

8. Davern, M. Social Networks and Economic Sociology: A Proposed Research Agenda for a More Complete Social Science [Text] / M. Davern // American Journal of Economics & Sociology. – 1997. – Vol. 56. – Is. 3. – P. 287–302.

9. Koren, Y. On Spectral Graph Drawing [Text] / Y. Koren // Proceedings of the 9th International Computing and Combinatorics Conference. – 2003. – P. 496–508.

10. Fortunato, S. Community Detection in Graphs [Text] / S. Fortunato // Phys. Rep. – 2010. – Vol.486. – No. 3–5. – P. 75–174.

11. Wasserman, S. Social Network Analysis: Methods And Applications [Text] / S. Wasserman, K. Faust. – N. Y.: Cambridge University Press, 1994. – 825 p.

12. Jensen, D. Data Mining in Social Networks [Text] / D. Jencen, J. Neville // Proceedings of the National Academy of

Sciences Symposium on Dynamic Social Network Analysis. – 2002. – P. 289–302.

13. Social Network Analysis and Mining for Business Applications [Text] / F. Bonchi, C. Castillo, A. Gionis, A. Jaimes // ACM TIST. – 2011. – Vol. 2 (3). – P. 22–58.

14. Hanneman, R. Computer-Assisted Theory Building: Modeling Dynamic Social Systems. Riverside [Text] / R. Hanneman. – CA: University of California, Riverside, 1988.

15. Leskovec, J. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations [Text] / J. Leskovec, J. Kleinberg, C. Faloutsos // Proceedings of the 11 th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD). – 2005. – P. 177–187.

16. Microscopic Evolution of Social Networks [Text] / J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins // Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2008. – P. 462–470.

17. Tantipathananandh, C. A Framework for Community Identification in Dynamic Social Networks [Text] / C. Tantipathananandh, T. Berger-Wolf, D. Kempe // Proceedings of the 13 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. –2007. – P. 717–726.

18. Graphscope: Parameter-Free Mining of Large Time-Evolving Graphs [Text] / J. Sun, C. Faloutsos, S. Papadimitriou, P. Yu // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2007. – P. 687–696.

19. Monitoring Network Evolution Using MDL [Text] / J. Ferlez, C. Faloutsos, J. Leskovec, D. Mladenic, M. Grobelnik // Proceedings of the International Conference on Data Engineering. – 2008. – P. 1328–1330.

20. Mining Graph Evolution Rules [Text] / M. Berlingerio, F. Bonchi, B. Bringmann, A. Gionis // Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science. – 2009. – Vol. 5781. – P. 115–130.

21. Desikan, P. Mining Temporally Changing Web Usage Graphs [Text] / P. Desikan, J. Srivastava // Proceedings of the International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles. 2004. P. 1–17.

22. Inokuchi, A. A Fast Method to Mine Frequent Subsequences from Graph Sequence Data [Text] / A. Inokuchi, T. Washio // Proceedings of the IEEE International Conference on Data Mining. – 2008. – P. 303–312.

23. Spotting Significant Changing Subgraphs in Evolving Graphs [Text] / Z. Liu, J. Yu, Y. Ke, X. Lin, L. Chen // Proceedings of the 8th International Conference on Data Mining. – 2008. – P. 917–922.

24. Borgwardt, K. M. Pattern Mining in Frequent Dynamic Subgraphs [Text] / K. M. Borgwardt, H.-P. Kriegel, P. Wackersreuther // Proceedings of the IEEE International Conference on Data Mining. – 2006. – P. 818–822.

25. Liben-Nowell, D. The Link Prediction Problem for Social Networks [Text] / D. Liben-Nowell, J. Kleinberg // Proceedings of the 12th International Conference on Information and Knowledge Management. – 2003. – P. 556–559.

26. Structure and Evolution of Blogspace / R. Kumar, J. Novak, P. Raghavan, A. Tomkins //Commun. ACM. – 2004. – Vol. 47. – No. 12. – P. 35–39.

27. Semantic Social Network Analysis [Text] / G. Érétéo, F. Gandon, M. Buffa, O. Corby // Proceedings of the 8th International Semantic Web Conference. – 2009. – P. 180–195.

28. Прохоров, А. Компьютерная визуализация социальных сетей [Text] / А. Прохоров, Н. Ларичев // КомпьютерПресс. – 2006. – № 9. – С. 156–160.

29. Huisman, M. A Reader's Guide to SNA Software [Text] / M. Huisman, Marijtje A. J. van Duijn // The SAGE Handbook of Social Network Analysis. SAGE. – 2011. – P. 578–600.

30. Adamic, L. A. The nature of markets in the World Wide Web [Text] / L. A. Adamic, B. A. Huberman // Quarterly Journal of Electronic Commerce 1, – 512 (2000).

31. Redner, S. How popular is your paper? [Text] / S. Redner // An empirical study of the citation distribution. Eur. Phys. J. – B 4. – 1998. – P.131–134.

32. Newman, M. E. J. Email networks and the spread of computer viruses [Text] / M. E. J. Newman, S. Forrest, J. Balthrop // Phys. Rev. E 66, 035101 (2002).

33. Simon, H. A. On a class of skew distribution functions [Text] / H. A. Simon // Biometrika 42. – 1955. – P. 425–440.

34. Shannon, C. E. A mathematical theory of communication [Text] / C. E. Shannon // I. Bell System Technical Journal 27. – 1948. – P. 379–423.

35. Yule, G. U. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis [Text] / G. U. Yule // Philos. Trans. R. Soc. London B 213. – 1925. – 21–87.

36. Erdős, P. On Random Graphs. I. [Text] / P. Erdős, A. Rényi // Publicationes Mathematicae 6. – 1959. – pp. 290-297.

37. Erdős, P. On the evolution of random graphs [Text] / P. Erdős, A. Rényi // Publ. Math. Inst. Hungar. Acad. Sci. 5. – 1960. – pp. 17-61.

38. Watts, D.J. Collective dynamics of "small-world" networks [Text] / D.J. Watts, S.H. Strogatz // Nature. – 1998. – Vol. 393. – pp. 440-442.

39. Albert, R. Attack and error tolerance of complex networks [Text] / R. Albert, H. Jeong, A. Barabasi // Nature. – 2000. – Vol. 406. – pp. 378-382.

40. Bjorneborn, L. Toward a basic framework for webometrics [Text] / L. Bjorneborn, P. Ingwersen // Journal of the American Society for Information Science and Technology. – 2004. – No 55(14). – pp. 1216-1227.

41. Vapnik, V.N. Statistical Learning Theory [Text] / V.N. Vapnik. –NY: John Wiley, 1998. – 760 p.

42. Manning, C.D. Foundations of Statistical Natural Language Processing [Text] / C.D. Manning, H. Schütze // Cambridge, Massachusetts: The MIT Press, 1999.

43. Reduction of English function words in Switchboard [Text] / A. Bell, E. Fosler-Lussier, C. Girand, W. Raymond // Proceedings of ICSLP-98. – Vol 7. – 1998. – pp. 3111-3114.

44. Salton, G. A Vector Space Model for Automatic Indexing [Text] / G. Salton, A. Wong, C. Yang // Communications of the ACM. – 1975. – 18(11) – 613-620.

45. Bradford, S.C. Sources of Information on Specific Subjects [Text] / S.C. Bradford // Engineering: An Illustrated Weekly Journal (London). – 1934. No. 137. – pp. 85-86.

46. Алексеев, Н. Г. Применение закона Бредфорда при комплектовании фонда научной библиотеки [Электронный ресурс] / Н. Г. Алексеев // Библиотечное дело-1996: тезисы докладов конференции. – Режим доступа: http://libconfs.narod.ru/1996/4s/4s_p1.html

47. Ландэ, Д. В. Основы интеграции информационных потоков [Электронный ресурс] / Д. В. Ландэ. – К.: Инжиниринг, 2006. – 240 с. – Режим доступа: http://dwl.visti.net/art/monogr-osnov/ spusk3.pdf

48. Grootjen, F.A. A formal derivation of Heaps' Law [Text] / F.A. Grootjen, D. C. Van Leijenhorst, van der T.P. Weide // Inf. Sci. – 2005. – Vol. 170(2-4). – pp. 263-272. – URL: http://citeseer.ist.psu.edu/660402.html

49. Barabási, A. Scale-free characteristics of random networks: the topology of the world-wide web [Text] / A. Barabási, R. Albert, H. Jeong // Physica. – 2000. – V. A281. – P. 69-77.

50. Barabási, L.-A. Diameter of the world-wide web [Text] / L.-A. Barabási, R. Albert, H. Jeong // Nature. – 1999. – V.401. – P. 130-131.

51. Buckley, P. G. Popularity basedrandom graph models leading to scale-free degree sequence [Text] / P. G. Buckley, D. Osthus // DiscreteMathematics. – 2004. –Vol. 282, Issues 1–3. – p.p. 53–68.

52. Directed scale-free graphs [Text] / B. Bollobás, C. Borgs, T. Chayes, O.M. Riordan // ProceedingSODA '03 Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms

53. Aiello, W. A Random Graph Model for Massive Graphs [Text] / W. Aiello, F. Chung, L. Lu // STOC'2000

54. Brin, S. The Anatomy of a Large-Scale Hypertextual Web Search Engine [Text] / S. Brin, L. Page // WWW7. – 1998.

55. Avram, H.D. The MARC II Format: A Communications Format for Bibliographic Data [Text] / H.D. Avram, J.F. Knapp, L.J. Rather // Library of Congress. – 1968.

56. Baeza-Yates, R. Modern Information Retrieval [Text] / R. Baeza-Yates, B. Ribeiro-Neto // N.Y.: ACM Press Series/Addison Wesley, 1999. – 513 p.

57. Арнольд, В. И. Аналитика и прогнозирование: математический аспект [Текст] / В. И. Арнольд // Научно-техническая информация. – Сер. 1. – Вып. 3. – 2003. – С. 1-10.

58. Heaps, H. S. Information Retrieval [Text] / H. S. Heaps // Computational and Theoretical Aspects. Academic Press, 1978.

59. Berry, M. W. Survey of Text Mining. Clustering, Classification, and Retrieval [Text] / M. W. Berry. – Springer-Verlag, 2004. – 244 p.

60. Kleinberg, J. M. Authoritative sources in a hyperlink environment [Text] / J. M. Kleinberg // In Processing of ACM-SIAM Symposium on Discrete Algorithms. – 1998. – 46(5). – pp. 604-632.

61. Hurst, H. E. Long-term storage capacity of reservoirs [Text] / H.E. Hurst // Trans. Amer. Soc. Civil Engineers 116. – 1951. – pp. 770-799.

62. Губанов, Д. А. Модели информационного влияния и информационного управления в социальных сетях [Текст] / Д. А. Губанов, Д. А. Новиков, А. Г. Чхартишвили // Проблемы управления. – 2009. – № 5. – 28–35.

63. Информационные риски в социальных сетях [Текст]: монография / Г. А. Остапенко, Л. В. Паринова, В. И. Белоножкин, И. Л. Батаронов, К. В. Симонов; под ред. чл.-корр. РАН Д. А. Новикова. – Воронеж: Издательство «Научная книга», 2013. – 160 с.

64. Жизнестойкость атакуемых распределенных систем: оценка рисков фатальных отказов компонентов [Текст]:

монография / А. Г. Остапенко, Д. Г. Плотников, О. Ю. Макаров, Н. М. Тихомиров, В. Г. Юрасов; под ред. чл.-корр. РАН Д. А. Новикова. – Воронеж: Издательство «Научная книга», 2013. – 160 с.

65. Управление информационными рисками мультисерверных систем при воздействии DDoS-атак [Текст]: монография / А. Е. Дешина, М. В. Бурса, А. Г. Остапенко, А. О. Калашников, Г. А. Остапенко; под ред. чл.-корр. РАН Д. А. Новикова. – Воронеж: Издательство «Научная книга», 2014. – 160 с.

66. Риск-анализ информационно-телекоммуникационных систем, подвергающихся атакам типа «сетевой шторм» [Текст]: монография/ А. Г. Остапенко, С. С. Куликов, Н. Н. Толстых, Ю. Г. Пастернак, Ю. Е. Дидюк; под ред. чл.-корр. РАН Д. А. Новикова. – Воронеж: Издательство «Научная книга», 2013. – 160 с.

67. Моделирование сложных атак на комплексные сети [Текст] / Ф. Галиндо, Н. В. Дмитриенко, А. Карузо, А. Россодивита, А. А. Тихомиров, А. И. Труфанов, Е. В. Шубников // Безопасность информационных технологий. – 2010. – № 3. – С.115-121.

68. Барабаши, А. Л. Безмасштабные сети [Текст] / А. Л. Барабаши, Э. Бонабо // В мире науки. – № 8. – 2003. – С. 55-63.

69. Mean-field Theory for Scale-free Random Network [Текст] / A. L. Barabashi, L. Yang-Yu, S. Jean-Jacques, R. Albert // Nature. – 2011. – No. 473– p. 167-173.

70. Albert, R. Statistical mechanics of complex networks [Текст] / R. Albert, A.L. Barabasi // Rev. Mod. Phys. 74 – 2002. – p. 47–97.

71. Assessment of the System’s Epi-resistanceUnder Conditions of Information Epidemic Expansion [Text] / N. M. Radko, A. G. Ostapenko, S. V.Mashin, O. A. Ostapenko, D. V. Gusev // Biosciences biotechnology research asia. – 2014. – Vol. 11(3). – 1781-1784.

72. Peak Risk Assessing The Process of Information Epidemics Expansion [Text] / N. M. Radko, A. G. Ostapenko, S. V.

Mashin, O. A. Ostapenko, A. S. Avdeev // Biosciences biotechnology research asia. – 2014. – Vol. 11(Spl. Edn.). – p. 251-255.

## ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ ..... 3
1. МЕТОДЫ АНАЛИЗА КОМПЬЮТЕРНЫХ СОЦИАЛЬНЫХ СЕТЕЙ ..... 5
1.1. Основные направления исследования компьютерных социальных сетей ..... 5
1.2. Некоторые наиболее известные социальные сети ..... 11
1.3. Параметры сложных сетей ..... 14
1.3.1. Параметры узлов сети ..... 14
1.3.2. Общие параметры сети ..... 15
1.3.3. Распределение степеней узлов ..... 15
1.3.4. Путь между узлами ..... 16
1.3.5. Коэффициент кластерности ..... 17
1.3.6. Посредничество ..... 18
1.3.7. Эластичность сети ..... 19
1.3.8. Структура сообщества ..... 19
1.4. Модели анализа социальных сетей ..... 20
1.5. Программные приложения для анализа социальных сетей ..... 39
2. ЭМПИРИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ И МАТЕМАТИЧЕСКИЙ ФОРМАЛИЗМ ..... 41
2.1. Эмпирические закономерности ..... 41
2.1.1. Распределение Парето ..... 46
2.1.2. Законы Ципфа ..... 49
2.1.3. Закономерность Бредфорда ..... 53
2.1.4. Закон Хипса ..... 54
2.2. Степенные распределения случайных величин ..... 55
2.3. Масштабно-инвариантные распределения ..... 58
2.4. Степенные законы для дискретных переменных ..... 60
3. МОДЕЛИ ФОРМИРОВАНИЯ И РОСТА СЕТЕЙ ..... 63
3.1. Модель Эрдеша-Реньи ..... 64
3.2. Наблюдения Барабаши-Альберт ..... 66
3.3. МодельLCD ..... 71
3.4. МодельBuckley-Osthus ..... 72
3.5. Модель копирования ..... 73

3.5.1. Геннерация графа ..... 73
3.5.2. Основной результат ..... 73
3.6. Ориентированные безмасштабные графы ..... 74
3.7. Модель Чунг-Лу ..... 74
3.7.1. Генерация графа ..... 74
3.7.2. Основные результаты ..... 75
3.7.3. Сравнение с реальными сетями ..... 75
3.8. Модель Янсона-Лучака ..... 75
3.8.1. Генерация графа ..... 75
3.8.2. Основные результаты ..... 76
3.8.3. Основные результаты для схожих моделей ..... 78
4. РАСПРОСТРАНЕНИЕ ЭПИДЕМИЙ В СОЦИАЛЬНЫХ ИНФОРМАЦИОННЫХ СЕТЯХ ..... 79
4.1. Модель эпидемии SI ..... 79
4.2. Модели просачивания и заражения ..... 82
4.3. Модель распространения эпидемии, адаптированная к социальным информационным сетям ..... 85
5. ИНФОРМАЦИОННЫЕ РИСКИ И ЭПИСТОЙКОСТЬ БЕЗМАСШТАБНЫХ СЕТЕЙ ..... 89
5.1. Риск-факторы безмасштабной сети ..... 94
5.2. Определение ущерба ..... 97
5.3. Оценка мощности ожидаемого множества успешно атакованных узлов ..... 105
5.4. Подходы к управлению эпистойкостью атакуемой безмасштабной сети ..... 111
ЗАКЛЮЧЕНИЕ ..... 114
БИБЛИОГРАФИЧЕСКИЙ СПИСОК ..... 115

Учебное издание

Шварцкопф Евгения Андреевна
Остапенко Ольга Александровна

# СОЦИАЛЬНЫЕ СЕТИ: РИСКИ И ОБЕСПЕЧЕНИЕ БЕЗОПАСНОСТИ

В авторской редакции

Подписано к изданию 27.04.2015.

Объем данных 1,27 Мб.

ФГБОУ ВПО «Воронежский государственный технический университет»
394026 Воронеж, Московский просп., 14