Моей самой важной социальной сети –
Сью, Алине и Эндрю – посвящается

Содержание

Предисловие ... 9

Глава 1. Введение в анализ сетей в R ... 11
1.1. Что такое сети? ... 12
1.2. Что такое анализ сетей? ... 14
1.3. Пять серьезных причин проводить анализ сетей в R ... 15
1.3.1. Широта возможностей R ... 15
1.3.2. Свободно распространяемая и открытая природа R ... 16
1.3.3. Возможности работы с данными и проектами в R ... 16
1.3.4. Широкий выбор пакетов для анализа сетей в R ... 17
1.3.5. Возможности моделирования сетей в R ... 17
1.4. Область применения книги и ресурсы ... 17
1.4.1. Область применения ... 17
1.4.2. «Дорожная карта» книги ... 18
1.4.3. Ресурсы ... 19

Часть I. ОСНОВЫ АНАЛИЗА СЕТЕЙ ... 20

Глава 2. «Пятичисловая сводка» для анализа сетей ... 21
2.1. Анализ в R: с чего начать ... 22
2.2. Подготовка ... 22
2.3. Простая визуализация ... 23
2.4. Базовое описание ... 23
2.4.1. Размер ... 23
2.4.2. Плотность ... 25
2.4.3. Компоненты ... 26
2.4.4. Диаметр ... 26
2.5. Коэффициент кластеризации ... 27

Глава 3. Управление сетевыми данными в R ... 28
3.1. Основные понятия сетевых данных ... 29
3.1.1. Структуры сетевых данных ... 29
3.1.2. Информация, хранимая в объектах-сетях ... 32
3.2. Создание объектов-сетей и работа с ними в R ... 32
3.2.1. Создание объекта-сети в `statnet` ... 33
3.2.2. Работа с атрибутами узлов и связей ... 36
3.2.3 Создание объекта-сети в `igraph` ... 39
3.2.4. Переключение между `statnet` и `igraph` ... 41
3.3. Импорт сетевых данных ... 41
3.4. Общераспространенные задачи при работе с сетевыми данными ... 43
3.4.1. Фильтрация сетевых данных на основе значений атрибутов вершин или ребер ... 43
3.4.2. Преобразование направленной сети в ненаправленную ... 50

Часть II. ВИЗУАЛИЗАЦИЯ ... 53

Глава 4. Графическое представление и укладка сети ... 54
4.1. Проблема визуализации сети ... 55
4.2. Эстетический вид укладок сетей ... 57
4.3. Основные алгоритмы и методы графического представления ... 59

4.3.1. Более точная настройка укладки сети ... 60
4.3.2. Укладки сетей, построенные с помощью igraph ... 62

Глава 5. Эффективный графический дизайн сетей ... 64
5.1. Основные принципы ... 65
5.2. Элементы дизайна ... 65
5.2.1. Цвет узла ... 66
5.2.2. Форма узла ... 71
5.2.3. Размер узла ... 72
5.2.4. Метка узла ... 77
5.2.5. Ширина ребра ... 78
5.2.6. Цвет ребра ... 79
5.2.7. Тип ребра ... 80
5.2.8. Легенды ... 81

Глава 6. Сложные графики сетей ... 83
6.1. Интерактивные графики сетей ... 84
6.1.1. Простые интерактивные сети в igraph ... 84
6.1.2. Публикация интерактивных веб-диаграмм сетей ... 85
6.1.3. Statnet Web: интерактивный statnet с помощью shiny ... 87
6.2. Специализированные диаграммы сетей ... 88
6.2.1. Дуговые диаграммы ... 88
6.2.2. Хордовые диаграммы ... 90
6.2.3. Теплокарты для сетевых данных ... 93
6.3. Создание диаграмм сетей с помощью других пакетов R ... 95
6.3.1. Построение диаграмм сетей с помощью ggplot2 ... 95

Часть III. ОПИСАНИЕ И АНАЛИЗ ... 99

Глава 7. Важность актора ... 100
7.1. Введение ... 101
7.2. Центральность – показатель важности для ненаправленных сетей ... 101
7.2.1. Три популярные меры центральности ... 103
7.2.2. Меры центральности в R ... 105
7.2.3. Централизация: вычисление индексов центральности для сети в целом ... 106
7.2.4. Создание отчетов по центральности ... 107
7.3. Точки сочленения и мосты ... 111

Глава 8. Подгруппы ... 114
8.1. Введение ... 115
8.2. Социальная сплоченность ... 116
8.2.1. Клики ... 116
8.2.2. k-ядра ... 120
8.3. Обнаружение сообществ ... 123
8.3.1. Модулярность ... 125
8.3.2. Алгоритмы обнаружения сообществ ... 127

Глава 9. Сети аффилированности ... 134
9.1. Определение сетей аффилированности ... 135
9.1.1. Аффилированность в виде бимодальных сетей ... 135
9.1.2. Двудольные графы (биграфы) ... 136
9.2. Основы сетей аффилированности ... 137
9.2.1. Создание сетей аффилированности из матриц инцидентности ... 137
9.2.2. Создание сетей аффилированности из списков ребер ... 138

9.2.3. Графическое представление сетей аффилированности ... 140
9.2.4. Проекции ... 140
9.3. Пример: актеры Голливуда как пример сети аффилированности ... 143
9.3.1. Анализ полной сети аффилированности актеров Голливуда ... 143
9.3.2. Анализ проекций актеров и фильмов ... 149

Часть IV. МОДЕЛИРОВАНИЕ ... 155

Глава 10. Модели случайных сетей ... 156
10.1. Предназначение моделей сетей ... 157
10.2. Модели формирования и структуры сети ... 158
10.2.1. Модель случайного графа Эрдеша–Реньи ... 158
10.2.2. Модель малого мира ... 162
10.2.3. Свободно масштабируемые модели ... 165
10.3. Сравнение моделей случайных графов с наблюдаемыми сетями ... 170

Глава 11. Статистические модели сетей ... 173
11.1. Введение ... 174
11.2. Построение экспоненциальных моделей случайных графов ... 177
11.2.1. Построение нулевой модели ... 179
11.2.2. Включение предикторов узлов ... 181
11.2.3. Включение предикторов диад ... 183
11.2.4. Включение предикторов ребер ... 187
11.2.5. Включение предикторов локальных структур (зависимых диадных связей) ... 189
11.3. Анализ экспоненциальных моделей случайных графов ... 191
11.3.1. Интерпретация модели ... 191
11.3.2. Подгонка модели ... 192
11.3.3. Диагностика модели ... 195
11.3.4. Имитационное моделирование сетей на основе оцененной модели ... 195

Глава 12. Модели динамических сетей ... 199
12.1. Введение ... 200
12.1.1. Динамические сети ... 200
12.1.2. RSiena ... 202
12.2. Подготовка данных ... 203
12.3. Спецификация и оценивание модели ... 210
12.3.1. Спецификация модели ... 210
12.3.2. Оценивание модели ... 214
12.4. Анализ модели ... 215
12.4.1. Интерпретация модели ... 215
12.4.2. Качество подгонки ... 220
12.4.3. Имитационное моделирование ... 224

Глава 13. Имитационные модели ... 228
13.1. Имитационные модели сетевой динамики ... 229
13.1.1. Имитационное моделирование социальной селекции ... 229
13.1.2. Имитационное моделирование социального влияния ... 240

Библиография ... 247

Предисловие

В начале 2000 года Стивен Хокинг сказал, что «следующий век будет веком сложности». Если его прогноз верен, то выходит, что нам потребуются новые научные теории, методы сбора данных и аналитические подходы, которые будут использоваться для исследования сложных систем и поведения. Наука о сетях – это подход, рассматривающий мир через призму сетей, в котором физические и социальные системы образованы разнородными акторами, соединенными друг с другом с помощью различных типов связей. Анализ сетей – это набор аналитических инструментов, используемых для изучения таких систем. В течение последних нескольких десятилетий анализ сетей приобретает все большее значение в арсенале аналитических средств, используемых социологами, врачами и физиками.

До недавнего времени для проведения анализа сетей требовалось специализированное программное обеспечение (как для управления сетевыми данными, так и для последующего анализа). Однако начиная примерно с 2000 года инструменты для анализа сетей появились в среде статистического программирования R. Помимо того что благодаря этому методы анализа сетей стали доступны более широкому кругу специалистов по статистике, пакет R предоставил исследователям, занимающимся анализом сетей, обширные возможности по управлению данными, графической визуализации и статистическому моделированию.

Как и предполагает название, эта книга является руководством пользователя по анализу сетей в R. В этой книге приводятся ключевые задачи в области анализа сетей, которые теперь можно выполнить в R. Книга концентрируется на четырех основных задачах, с которыми обычно сталкивается специалист в области анализа сетей: управление сетевыми данными, визуализация сети, описание сети и моделирование сети. Книга включает программный код R, который используется в конкретных примерах анализа сетей. Кроме того, к книге прилагается комплект наборов сетевых данных, использующихся в ней. (См. главу 1 для получения более подробной информации о структуре книги, а также инструкции по поводу того, как получить сетевые данные.) Книга написана для тех, кого интересует проведение анализа сетей в R. Она может использоваться в качестве вспомогательного пособия по анализу сетей или руководства по методам анализа сетей в R.

Появление этой книги было бы невозможным без консультаций, поддержки, рекомендаций и советов, которые я получил за последние 30 лет благодаря своим собственным социальным сетям (личной и профессиональной). В середине 1980-х годов я закончил класс по анализу сетей у Стена Вассермана в Иллинойском университете в Урбане-Шампейне. Я помню, в каком я был восторге от этого нового метода анализа данных, но тогда думал, что вряд ли буду когда-либо использовать его в своей работе. Однако мои коллеги в области психологии и здравоохранения посоветовали мне в моей первоначальной работе рассмотреть тему использования анализа сетей для изучения и оценки данных. Среди них – Джулиан Раппапорт (Julian Rappaport), Эд Сейдман (Ed Seidman), Брюс Рапкин (Bruce Rapkin), Курт Рибисл (Kurt Ribisl), Шерон Хоман (Sharon Homan), Росс Браун-

сон (Ross Brownson) и Мэтт Кройтор (Matt Kreuter). Независимо от того, знают они это или нет, я был вдохновлен замечательным коллективом специалистов в области сетей и систем, в их числе Том Валенте (Tom Valente), Стив Боргатти (Steve Borgatti), Мартина Моррис (Martina Morris), Том Снайдерс (Tom Snijders), Скотт Лейшоу (Scott Leischow), Пэтти Мейбри (Patty Mabry), Стивен Маркус (Stephen Marcus) и Росс Хаммонд (Ross Hammond). Свои главные идеи, связанные с анализом сетей, я почерпнул от моих друзей и коллег в Научном центре общественного здравоохранения, в частности от Бобби Карозерса (Bobbi Carothers), Амара Дхенда (Amar Dhand), Криса Робишо (Chris Robichaux) и Нэнси Мюллер (Nancy Mueller). Особенно я благодарен моим студентам, посещавшим мои занятия и семинары на протяжении этих лет. Они не только улучшили эту книгу, но и расширили мои взгляды касательно анализа сетей. Отдельное большое спасибо Дженин Харрис (Jenine Harris). Дженин была моим первым докторантом, и в данный момент я восхищен строгостью и элегантностью ее работы, посвященной анализу сетей. Я также хотел бы поблагодарить Центры по контролю и профилактике заболеваний США, Национальные институты здравоохранения США и Фонд здоровья в Миссури за поддержку в проведении исследований, что позволило мне разработать и усовершенствовать подход к анализу сетей. Наконец, выражаю глубочайшую признательность членам моей семьи. Они дали мне определенные советы касательно содержания, предоставили место и время для напряженной работы над этой книгой (включая знаменательный подарок на День отца) и поддерживали меня в те моменты, когда я больше всего в этом нуждался. Спасибо вам, Сью, Али и Эндрю.

Сент-Луис, Миссури, США Дуглас Люк
Июль 2015

Глава 1

Введение в анализ сетей в R

1.1. Что такое сети? 12
1.2. Что такое анализ сетей? 14
1.3. Пять серьезных причин проводить анализ сетей в R 15
1.4. Область применения книги и ресурсы 17

– Начните сначала, – серьезно сказал Король, – и читайте, пока не дойдете до конца: тогда и остановитесь.

Льюис Кэрролл. «*Алиса в Стране чудес*»

1.1. Что такое сети?

Эта книга является руководством пользователя для проведения анализа сетей в среде статистического программирования R. Сети – это все, что окружает нас. Люди естественным образом организуются в сетевые системы. Наши близкие и друзья формируют персональные социальные сети вокруг каждого из нас. Соседские общины организуются в сетевые объединения для выдвижения тех или иных требований. Компании сотрудничают (или конкурируют) друг с другом в рамках сложных, взаимосвязанных отношений торгового и финансового партнерства. Развитие здравоохранения осуществляется путем партнерства правительственных и неправительственных организаций [Luke, Harris, 2007]. Страны связаны друг с другом системами миграции, торговли и договорных обязательств.

Кроме того, практически везде встречаются сети, не связанные с человеческими коммуникациями. Наши гены и белки взаимодействуют друг с другом посредством сложных биологических сетей. Человеческий мозг теперь рассматривается как сложная сеть, или «коннектом» («connectome») [Sporns, 2012]. Аналогично человеческие болезни и их базовые генетические корни можно представить в виде «карты болезни» («diseasome») [Barabasi, 2007]. Виды животных взаимодействуют друг с другом различными сложными способами, один из которых – пищевая сеть, в которой взаимодействия можно описать отношениями «кто кого съедает». Информация уже сама по себе объединена в сеть. Наша правовая система представляет собой взаимосвязанную сеть ранее принятых юридических решений и прецедентов. Социальный и научный прогресс стимулируется процессом распространения инноваций, в ходе которого информация разносится по взаимосвязанным социальным системам, будь то фермеры Айовы [Rogers, 2003] или специалисты в области общественного здравоохранения [Harris, Luke, 2009]. Похоже, что сети являются одним из способов, с помощью которого устроена вселенная.

Так что же такое сеть? На рис. 1.1 и 1.2 показаны две важные и интересные социальные сети. Рисунок 1.1 представляет собой сеть контактов 19 налетчиков, совершивших террористическую атаку на США 11 сентября 2001 года. Она взята из работы [Valdis Krebs, 2002]. Социальная сеть состоит из множества акторов (также называемых узлами), которые соединены друг с другом определенным типом социальных отношений (также называемых связью).

На рисунке узлы показаны кружками, а связи – это линии, соединяющие некоторые узлы. Сеть показывает нам, что налетчики контактировали друг с другом, прежде чем совершить теракт 11 сентября, но количество связей в сети небольшое и, кажется, нет никакого доминирующего участника сети, который был бы связан со всеми налетчиками или с большинством из них.

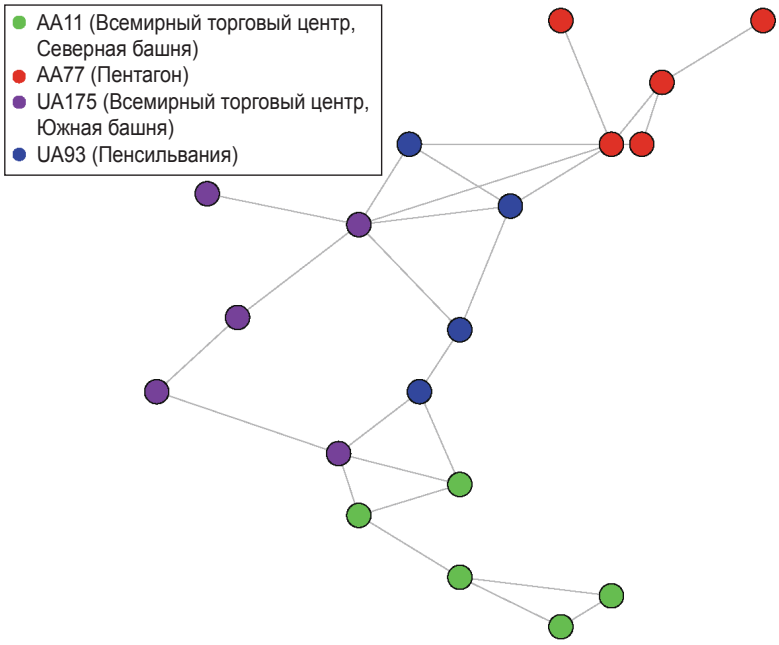

Рис. 1.1. Сеть контактов налетчиков, совершивших террористическую атаку на США 11 сентября 2001 года

Второй пример, приведенный на рис. 1.2, представляет совсем другой вид социальной сети. Здесь узлы – это участники сборной Нидерландов на Чемпионате мира по футболу FIFA 2010 года, которая потерпела поражение в финальном матче с Испанией. Связи – это передачи мяча между различными игроками во время матчей на Чемпионате мира. Стрелки показывают направление передач. Мы видим, что вратарь передавал мяч прежде всего защитникам, а нападающие получали пасы главным образом от полузащитников (за исключением игрока под номером 6, который, в отличие от двух остальных нападающих, похоже, использовал другую манеру передачи мяча).

Может показаться, что между этими двумя примерами мало общего. Однако их объединяет фундаментальное свойство, характерное для всех социальных сетей. Типы социальных взаимосвязей, приведенные на рисунках, не случайны. Они отражают лежащие в их основе социальные процессы, которые можно исследовать с помощью научных теорий и методов, использующихся в рамках анализа сетей. Террористическая сеть не имеет выраженного лидера, и ее члены слабо связаны друг с другом, поскольку именно это и затрудняет обнаружение и ликвидацию такой сети. Связи, представляющие собой паттерны передачи мяча, отражают роли игроков, правила игры и стратегию тренера. Анализ сетей «понятия не имеет» об этих правилах или стратегиях. И тем не менее его можно использовать для определения этих паттернов, которые отражают основные правила и закономерности.

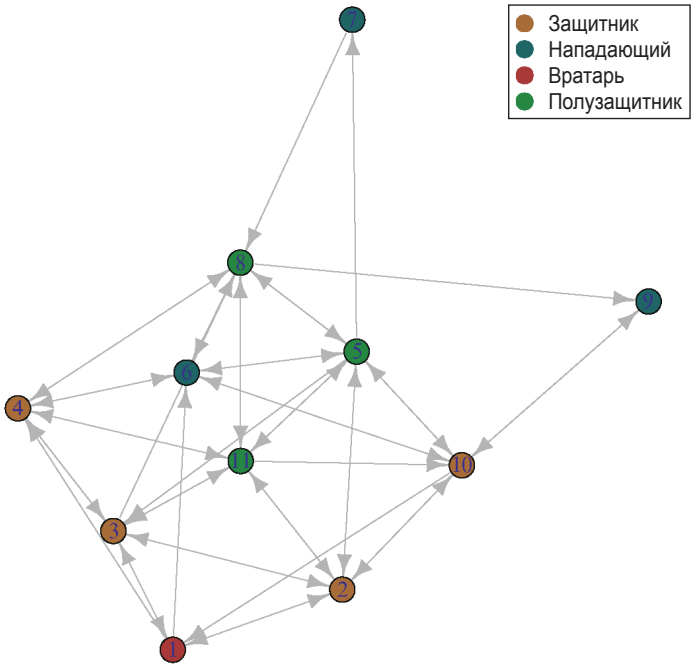

Рис. 1.2. Сеть игроков сборной Нидерландов, участвовавшей в Чемпионате мира по футболу FIFA 2010 года

1.2. Что такое анализ сетей?

Наука о сетях (*network science*) является широким научным подходом, который использует анализ взаимосвязей для изучения и интерпретации биологических, физических, социальных и информационных систем. Основной инструмент для специалистов по сетям – это анализ сетей, который представляет собой набор методов, использующихся для (1) визуализации сетей, (2) описания определенных характеристик структуры сети в целом, а также получения детальной информации об отдельных узлах, связях и подгруппах внутри сетей, (3) создания математических и статистических моделей сетевых структур и сетевой динамики. Поскольку основной вопрос, интересующий науку о сетях, касается взаимосвязей, большинство методов, используемых в анализе сетей, сильно отличается от более традиционных статистических инструментов, используемых социологами и учеными-медиками.

Анализ сетей (*network analysis*) – это отдельное научное направление со своими собственными теориями и методами, взятыми из других дисциплин, в частности из теории графов и топологии в математике, анализа систем родственных связей в антропологии, анализа социальных групп и процессов в социологии и психологии. Хотя анализ сетей не был изобретен конкретным человеком в определенном месте и в определенное время, первоначальные наработки того, что мы теперь

называем современным анализом сетей, можно найти в работе Якоба Морено , опубликованной в 1930-х годах. Он стал называть исследования социальных отношений социометрией и основал журнал *Sociometry*, в котором были опубликованы первые работы в этой области. Он также изобрел социограмму, которая была средством визуализации структуры сети. Первая опубликованная социограмма появилась в газете «Нью-Йорк таймс» в 1933 году, и это была сетевая диаграмма дружеских отношений между учениками 4-го класса. (Эта информация входит в набор сетевых данных, который используется в данной книге, см. раздел 1.4.3 ниже.)

Теории и методы анализа сетей разрабатывались на протяжении последних десятилетий XX века, существенный вклад внесли социология, психология, политология, бизнес, здравоохранение и компьютерная наука. Развитие науки о сетях в качестве практической дисциплины было обусловлено разработкой различных программных средств для анализа сетей, включая UCINet, STRUCTURE, Negopy и Pajek. В течение последних 20–30 лет отмечается взрывной рост популярности науки о сетях, обусловленный, по крайней мере, тремя различными факторами. Во-первых, математики, физики и другие исследователи разработали много серьезных теорий, касающихся формирования сетей и их структур, что привлекло внимание к науке о сетях (см. главу 10, в которой рассматриваются эти теории). Во-вторых, рост вычислительной мощности и скорости позволил использовать методы анализа сетей применительно к большим и очень большим сетям, таким как Интернет, население планеты или человеческий мозг. Наконец, новые подходы, появившиеся в рамках статистической теории сетей, впервые позволили аналитикам выйти за рамки простого описания сети, чтобы построить и протестировать статистические модели сетевых структур и процессов (см. главы 11 и 12).

1.3. Пять серьезных причин проводить анализ сетей в R

Как и предполагает название, эта книга является общим руководством для проведения анализа сетей с помощью статистического языка и программной среды R. Почему R является идеальной платформой для разработки и проведения анализа сетей? Существует, по крайней мере, пять серьезных доводов.

1.3.1. Широта возможностей R

Язык и среда статистического программирования R представляют собой обширную интегрированную систему, состоящую из тысяч пакетов и функций, которые позволяют выполнять неисчислимое количество задач по управлению данными, анализу и визуализации. R включает в себя ряд пакетов, которые предназначены для выполнения определенных задач анализа сетей. Выполняя эти задачи в среде R, аналитик может использовать в своих интересах все имеющиеся возможности R. Большинство остальных программ по анализу сетей (например, Pajek, UCINet, Gephi) являются автономными пакетами и поэтому не обладают преимуществами работы в интегрированной среде статистического программирования.

1.3.2. Свободно распространяемая и открытая природа R

Одной из важных причин популярности и успеха R является тот факт, что R – это свободно распространяемый и открытый язык. Официально это подкреплено Открытым лицензионным соглашением GNU, в соответствии с которым распространяется программный код R. Если говорить менее формально, существует огромное сообщество пользователей и разработчиков R, которые постоянно работают над расширением возможностей и улучшением программного кода R и тысячи пакетов R, доступных для свободного скачивания. Инструменты R по анализу социальных сетей, описанные в этой книге, были фактически разработаны пользовательским сообществом R. Открытая природа R способствует более быстрой (и возможно, более внятной и более функциональной) разработке и популяризации новых статистических техник, например инструментов для анализа сетей.

1.3.3. Возможности работы с данными и проектами в R

Несмотря на то что существует много хороших общедоступных программ по анализу сетей, которые могут выполнить самые различные задачи, связанные с исследованием описательных статистик сетей и визуализацией, ни один из программных пакетов, по сравнению с R, не имеет таких же возможностей работы с данными и проектами по изучению крупномасштабных сетей. Во-первых, как уже упоминалось выше, для анализа сетей в R можно использовать внушительные возможности языка R по управлению, очистке, импорту и экспорту данных. Как описано в главе 3, анализ сетей часто начинается с того, что нужно импортовать данные из других источников и преобразовать их в формат, который можно проанализировать с помощью соответствующих инструментов. Все программные пакеты по анализу сетей имеют определенные инструменты управления данными, но ни одна из программ не сравнится с R по широте возможностей и глубине анализа.

Во-вторых, при проведении сложного научного или прикладного анализа сетей важно иметь под рукой подходящие инструменты управления проектами, чтобы упростить хранение и извлечение программного кода, работу с аналитическими выводами (например, со статистическими результатами и графиками), составление отчетов для внутренних и внешних пользователей. В отличие от большинства программ по работе с сетями, традиционные платформы статистического анализа, например SAS и SPSS, имеют подобные инструменты. Используя интегрированную среду разработки для языка R, например RStudio (http://rstudio.org/), и вооружившись преимуществами таких пакетов, как `knitr` и `shiny`, пользователь получает возможность управлять проектом сети любой сложности. Фактически легкость разработки и общедоступность этих инструментов были движущими силами направления *воспроизводимое исследование* (*reproducible research*) [Gentleman, Lang, 2007], которое подчеркивает важность объединения данных, программного кода, результатов и документации в унифицированном и общедоступном формате. В качестве примера, иллюстрирующего возможности инструментов воспроизводимого исследования, имеющихся в R, можно привести эту книгу, которая полностью была создана в RStudio.

1.3.4. Широкий выбор пакетов для анализа сетей в R

Основная причина, по которой R идеально подходит для анализа сетей, – это широкий выбор пакетов, которые в данный момент позволяют управлять сетевыми данными, а также осуществлять визуализацию, интерпретацию и моделирование сетей. Существуют десятки пакетов для анализа сетей, и их становится все больше и больше. Благодаря пакетам `network` и `igraph` с сетевыми данными можно работать как с объектами R, с помощью пакета `intergraph` объекты одного класса можно преобразовать в объекты другого класса. Базовый анализ и визуализацию сетей можно выполнить с помощью пакета `sna`, который входит в более широкий комплект пакетов для анализа сетей `statnet`, а также в `igraph`. Более сложное моделирование сетей можно выполнить с помощью пакета `ergm` и сопутствующих библиотек, модели динамических сетей строятся с помощью пакета `RSiena`. Свободно распространяемые программы анализа сетей имеют много преимуществ (здесь можно упомянуть о возможностях визуализации в Gephi), но ни одна из программ не сравнится с объединенными возможностями пакетов для анализа социальных сетей, имеющихся в R.

1.3.5. Возможности моделирования сетей в R

Наконец, нужно упомянуть о конкретных преимуществах R с точки зрения моделирования сетей. R является единственным общедоступным программным пакетом, который обладает всесторонними возможностями для проведения стохастического моделирования сетей (например, в R можно создавать экспоненциальные модели случайных графов), построения моделей динамических сетей (что позволяет исследовать изменение сети с течением времени) и имитационного моделирования сетей.

1.4. Область применения книги и ресурсы

1.4.1. Область применения

Как следует из названия, цель этой книги состоит в том, чтобы дать вам практическое руководство по анализу сетей в среде статистического программирования R. Книга является практической в том смысле, что в ней приводятся короткие фрагменты программного кода, которые предназначены для проведения анализа сетей и применяются к реальным сетевым данным. Здесь же приводятся результаты исследований. Читатель может скачать примеры программного кода и данные, чтобы легко повторить все то, что показано в книге, поэкспериментировать с собственными данными или программным кодом и таким образом упростить процесс обучения.

Практическая цель книги состоит в том, чтобы продемонстрировать применение методов анализа сетей в R для решения различных исследовательских задач. Речь идет об управлении данными, визуализации сети, вычислении описательных статистик сети и выполнении математического, статистического и динамичного моделирования сетей. Целевая аудитория включает студентов, аналитиков и исследователей различного профиля, конкретно социологов, ученых-медиков, представителей бизнеса и инженеров.

Кроме того, необходимо отметить, что эта книга не является инструкцией по выполнению анализа сетей. Во-первых, в этой книге нет всестороннего рассмотрения теорий, развивавшихся в рамках науки о сетях. Существует много хороших книг, статей, учебных курсов и интернет-ресурсов в свободном доступе, где излагается этот материал. Для составления общего представления все еще подойдет классический текст [Wasserman, Faust, 1994], а в работе [John Scott, 2012] представлен прекрасный и более актуальный обзор теорий и методов науки о сетях. Для более глубокого знакомства с наукой о сетях и статистической теорией см. работу [Newman, 2010] или [Kolaczyk, 2009]. Наконец, упомяну о двух работах, в которых подробно освещена новейшая история науки о сетях, а также прекрасно реализованы научные исследования эмпирических сетей, – работе [Newman et al., 2006], а также работе [Scott, Carrington, 2011].

Во-вторых, эта книга ни в коем случае не является введением в программирование R и статистический анализ. Несмотря на то что были предприняты все попытки сделать примеры программного кода максимально понятными и краткими, начинающий пользователь R обнаружит, что некоторые приемы и синтаксис программного кода сложно понять. В частности, чтобы извлечь максимальную пользу из этого руководства, вам очень пригодится знакомство с возможностями R по управлению данными, графикой и объектно-ориентированным подходом к статическому моделированию.

Таким образом, книга адресована заинтересованному студенту, аналитику или исследователю, который знаком с R и имеет некоторое представление о теории и методах науки о сетях. Ее можно использовать в качестве вспомогательного пособия по анализу сетей для студентов вузов. Кроме того, опытные аналитики, работающие в R и желающие включить анализ сетей в свой арсенал программно-аналитических средств, могут использовать эту книгу в качестве учебника.

1.4.2. «Дорожная карта» книги

Книга разбита на четыре основные части, которые соответствуют четырем фундаментальным задачам, на выполнение которых специалисты по анализу сетей тратят бóльшую часть своего времени: управление данными, визуализация сетей, описание сетей и моделирование сетей. Первая часть включает две главы, которые представляют собой введение в базовые методы анализа сетей, затем следует более глубокое рассмотрение проблем, связанных с управлением данными в рамках анализа сетей. Три главы, относящиеся к части «Визуализация», посвящены базовым укладкам сетей, выбору графического дизайна сетей, также рассматриваются некоторые продвинутые графики и методы визуализации. Часть «Описание и анализ» состоит из трех глав, которые касаются наиболее распространенных методов, используемых для описания важных характеристик сети, включая важность актора, подгруппы и сообщества внутри сетей, работу с сетями аффилированности. Заключительная часть «Моделирование» включает четыре главы, которые представляют собой продвинутые методы математического моделирования, статистического моделирования, моделирования динамических сетей и имитационного моделирования сетей. В табл. 1.1 приводится «дорожная карта» книги.

Таблица 1.1. «Дорожная карта» руководства пользователя

Глава	Пакеты	Наборы данных
Введение		FIFA_Nether, Krebs
Пятичисловая сводка	statnet, sna	Moreno
Сетевые данные	statnet, network, igraph	DHHS, ICTS
Базовая визуализация	statnet, sna	Moreno, Bali
Графический дизайн	statnet, sna, igraph	Bali
Продвинутая графика	arcdiagram, circlize, visNetwork, networkD3	Simpsons, Bali
Важность	statnet, sna	DHHS, Bali
Подгруппы	igraph	DHHS, Moreno, Bali
Сети аффилированности	igraph	hwd
Математические модели	igraph	lhds
Стохастические модели	ergm	TCnetworks
Динамические модели	RSiena	Coevolve
Имитационное моделирование	igraph	

1.4.3. Ресурсы

Важнейшим ресурсом для этого руководства является коллекция наборов сетевых данных, специально подобранных и доступных для скачивания. Более десятка наборов сетевых данных включены в пакет R под названием `UserNetR`. Эти наборы данных используются на протяжении всей книги для иллюстрации примеров программного кода и анализа. Наборы сетевых данных, включенные в пакет `UserNetR`, преимущественно взяты из опубликованных исследований по анализу сетей, а некоторые созданы специально, чтобы проиллюстрировать решение конкретных аналитических задач. В табл. 1.1 перечисляются названия наборов данных, которые используются в каждой главе.

Пакет `UserNetR` находится на GitHub. Чтобы получить доступ к сетевым данным, нужно скачать и установить этот пакет. Это можно сделать с помощью программного кода, приведенного ниже. Кроме того, нужно установить пакет `devtools`, если он у вас отсутствует.

```
library(devtools)
install_github("DougLuke/UserNetR")
```

Как только это сделано, нужно загрузить `UserNetR`, чтобы получить доступ к различным файлам данных. Как и любой пакет R, его можно загрузить с помощью функции `library()`. Эта команда не всегда приводится в книге, поэтому, прежде чем выполнить любой приведенный программный код R, удостоверьтесь в том, что загрузили пакет `UserNetR`.

```
library(UserNetR)
```

С документацией по пакету `UserNetR` можно ознакомиться с помощью справочной системы R.

```
help(package='UserNetR')
```

ЧАСТЬ I

ОСНОВЫ АНАЛИЗА СЕТЕЙ

Глава 2. «Пятичисловая сводка» для анализа сетей 21
Глава 3. Управление сетевыми данными в R 28

Глава 2

«Пятичисловая сводка» для анализа сетей

2.1. Анализ в R: с чего начать.....22
2.2. Подготовка 22
2.3. Простая визуализация 23
2.4. Базовое описание.....23
2.5. Коэффициент кластеризации ...27

Когда ищешь, то обязательно находишь. Спору нет, если ищешь, то всегда что-нибудь найдешь, но совсем не обязательно то, что искал.

Д. Р. Р. Толкин. «*Хоббит*»

2.1. Анализ в R: с чего начать

С чего нужно начать анализ сетей в R? Разумеется, ответ на этот вопрос зависит от аналитических задач, которые вы надеетесь решить, состояния имеющихся у вас сетевых данных и целевой аудитории, которой будут адресованы результаты этого анализа. Хорошей новостью, связанной с выполнением анализа сетей в R, является тот факт (проиллюстрированный в последующих главах), что R предлагает множество инструментов для проведения анализа сетей. Однако принять решение, с какого инструмента начать, – непростая задача.

В 1977 году Джон Тьюки предложил пятичисловую сводку[1] в качестве простого и быстрого способа кратко описать самые важные характеристики одномерного распределения. Сети сложнее отдельных переменных, однако можно провести анализ важных характеристик социальной сети, воспользовавшись небольшим количеством процедур в R.

В этой главе мы сфокусируемся на двух первоначальных шагах, без которых невозможно приступить к анализу сетей: простой визуализации и базовом описании сети с помощью «пятичисловой сводки». Кроме того, эта глава является плавным введением в основы анализа сетей в R и рассказывает о том, как можно быстро выполнить этот анализ.

2.2. Подготовка

Как и в других видах статистического анализа, выполняемого в R, сначала необходимо загрузить соответствующие пакеты (предварительно установив их, если это необходимо) и получить доступ к данным. В данном случае для анализа будет использоваться комплект пакетов `statnet`. Данные, использованные в этой главе (и в остальной части книги), взяты из пакета `UserNetR`, который будет использоваться на протяжении всей книги. Конкретный набор данных, использованный здесь, называется `Moreno` и содержит сеть дружеских отношений между учениками 4-го класса, впервые составленную Якобом Морено в 1930-х годах.

```
library(statnet)
library(UserNetR)
data(Moreno)
```

[1] Пятичисловая сводка Тьюки включала медиану (Q2), первый (Q1) и третий (Q3) квартили, наименьшее (min) и наибольшее (max) значения. – *Прим. пер.*

2.3. Простая визуализация

Первый шаг в анализе сетей часто заключается в том, чтобы просто посмотреть на сеть. Визуализация сети весьма важна, но, как следует из глав 4, 5 и 6, построение эффективного графика сети требует тщательного планирования и реализации. Вместе с тем информативный график сети можно построить с помощью одного простого вызова функции. Единственная дополнительная сложность заключается в том, что цвет узла определяется полом участника сети. Детали синтаксиса, лежащего в основе этого примера, будут более подробно рассмотрены в главах 3, 4 и 5.

```
gender <- Moreno %v% "gender"
plot(Moreno, vertex.col = gender + 2, vertex.cex = 1.2)
```

Получившийся график четко показывает, что сеть дружеских контактов состоит из двух довольно четких подгрупп, образованных по половому признаку. Оперативно построенный график сети, аналогичный этому, часто показывает важнейшие структурные паттерны социальной сети.

2.4. Базовое описание

Исходная пятичисловая сводка Тьюки была предназначена для описания важнейших характеристик распределения переменной (включая центральную тенденцию и изменчивость) с помощью статистических показателей. Аналогично, используя лишь несколько функций и строк программного кода R, мы можем построить пятичисловую сводку для сети, которая ответит нам на следующие вопросы: каков *размер* сети, какова *плотность* сети, состоит ли сеть из одной или нескольких отдельных *групп*, насколько сеть *компактна* и как *кластеризованы* участники сети.

2.4.1. Размер

Самая главная характеристика сети – это ее *размер* (*size*). Размер – это просто количество *участников* (*members*), обычно называемых *узлами* (*nodes*), *вершинами* (*vertices*) или *акторами* (*actors*). Самый простой способ получить информацию о размере – воспользоваться функцией `network.size()`. Базовая сводка для объекта-сети `statnet`, полученная с помощью функции `summary`, также предоставляет эту информацию.

Благодаря `network.size` и `summary` узнаем, что сеть Moreno состоит из 33 участников. (Значение `false`, установленное для `print.adj`, подавляет вывод подробной информации о *смежности* (*adjacency*), который займет много места.)

```
network.size(Moreno)

  ## [1] 33

summary(Moreno,print.adj=FALSE)

  ## Network attributes:
  ##   vertices = 33
```

```
##   directed = FALSE
##   hyper = FALSE
##   loops = FALSE
##   multiple = FALSE
##   bipartite = FALSE
##  total edges = 46
##    missing edges = 0
##    non-missing edges = 46
##  density = 0.0871
##
## Vertex attributes:
##
##   gender:
##     numeric valued attribute
##     attribute summary:
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.00    1.00    2.00    1.52    2.00    2.00
##    vertex.names:
##     character valued attribute
##     33 valid vertex names
##
## No edge attributes
```

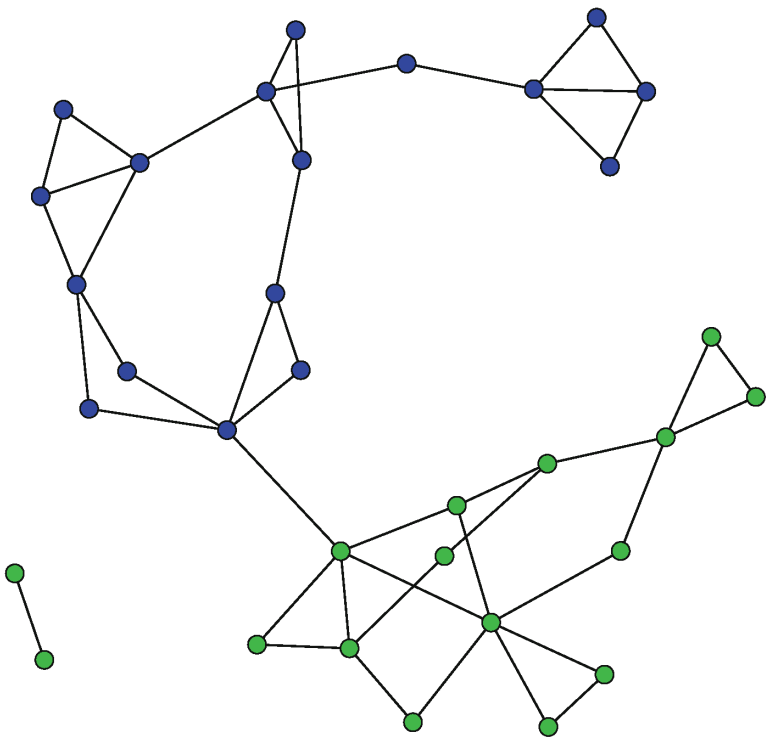

Рис. 2.1. Социограмма Морено