МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра дискретной математики и информационных технологий

ПРИМЕНЕНИЕ ЭНТРОПИЙНЫХ МЕР ДЛЯ АНАЛИЗА СЛОЖНЫХ СЕТЕЙ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Алексаненковой Екатерины Павловны

Научный руководитель
доцент, к. ф.-м. н. _____ И. Д. Сагаева

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2020

ВВЕДЕНИЕ

Актуальность работы. Многие реальные сети представляют собой сложные сети, такие как социальные, информационные, технологические, финансовые и биологические сети. В последнее время сложные сети всё больше привлекают внимание исследователей из разных областей, таких как физика, математика, биология, медицина, инженерия и компьютерные науки. Многочисленные исследования показали, что свойства структуры играют важную роль в исследовании сложных сетей.

Сложные сети способны моделировать самые разнообразные структуры, поддерживающие функционирование повседневной жизни, например, – интернет, коммуникационные, химические, нейронные, социальные, политические и финансовые сети.

Данная тема актуальна из-за большого числа областей применения. В частности, сложные сети особенно важны в области современных социальных сетей.

Количественная оценка сложности сетей на сегодняшний день является фундаментальной проблемой в физике сложных систем. Возможное решение проблемы – применение понятий теории информации в сетях. В данной работе для анализа сетей используется квантификатор теории информации такой как энтропия сети.

Цель магистерской работы – исследование применения мер энтропии для анализа сложных сетей.

В соответствии с поставленной целью определены следующие задачи:

1. Изучить существующие меры энтропии, которые используются при измерении сложных сетей и графов;
2. Изучить внутренние и внешние модели оценки энтропии;
3. Провести анализ области применения энтропии;
4. Изучить природу энтропии в моделях Эрдёша–Реньи и Барабаши–Альберт;
5. Оценить возможность применения энтропийных мер для предупреждения системных рисков;
6. Оценить возможность применения энтропийных мер в онлайн социальной сети;
7. Разработать программный код для расчета мер энтропии как в модель-

ных графах, так и в реальных сетях.

Методологические основы исследования энтропийных мер и анализа сложных сетей представлены в работах Эрдёша-Реньи, Барабаши-Альберт и Бьянкони.

Практическая значимость магистерской работы. Компьютерные сети становятся все более важными в современной информатике так как они предоставляют целостную модель для представления многих реальных явлений. Обилие данных о взаимодействиях в сложных системах позволяет сетевым наукам описывать, моделировать и прогнозировать поведение и состояния сложных систем. Поэтому важно охарактеризовать сети с точки зрения их сложности, чтобы приспособить аналитические методы к конкретным сетям. Мера сложности сети необходима для многочисленных приложений. Например, уровень сложности сети может определять ход различных процессов, происходящих в сети, таких как распространение информации, распространение сбоев, действия, связанные с управлением, или сохранение устойчивости.

Структура и объём работы. Магистерская работа состоит из введения, 5 разделов, заключения, списка использованных источников и 4 приложений. Общий объем работы – 73 страницы, из них 59 страниц – основное содержание, включая 16 рисунков и 6 таблиц, список использованных источников информации – 25 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Анализ литературы» посвящён изучению литературы, касающейся тематики данной ВКР, а именно современных публикаций на такие темы как «модели случайных графов», «энтропия сети» и «меры энтропии».

Рассматриваются модели случайных графов Эрдёша-Реньи, Барабаши-Альберт, Боллобаша-Риордана и их применения. Так же в рамках данного раздела было изучено понятие *энтропия сети.* Изучены энтропии Шеннона, Гиббса и Корнера. Определены простые формулы для определения энтропии определённых семейств классов графов.

Второй раздел «Анализ области применения» посвящён анализу областей применения мер энтропии – финансовая система и социальная

сеть. Рассматриваются только те две области применения энтропийных мер, которые были задействаны в ВКР.

Энтропийные меры могут применяться для раннего предупреждения системных рисков в финансовой системе. Мотивация исследования именно в этой сфере основывается на способности обнаруживать и прогнозировать сходство системных событий, определяемых как состояние финансового стресса. В связи с этим предлагается подход, позволяющий фиксировать структурные изменения системы. Идея заключается в том, что системный риск связан с общей изменчивостью в единую финансовую систему, где существуют связи между учреждениями, например банки, страховые компании, финансовые дилеры – это условия для наблюдения каскадных эффектов потерь. То есть, если учреждение испытывает состояние бедствия, то другие учреждения, связанные с первым, также могут пострадать.

Вторая область применения, рассматриваемая в магистерской работе – онлайн социальная сеть *Twitter*. *Twitter* используется по разным причинам, включая распространение информации, маркетинг, распространение пропаганды, рассылки спама, для общения и так далее. Охарактеризовать эти действия и классифицировать связанный с ними пользовательский контент является сложной задачей. Классифицировать данную пользовательскую активность в *Twitter* можно с помощью методов сетевой энтропии. Для этого нужно изучить их общую динамику «ретвитов». Для классификации ретвитов необходимы две функции: временной интервал и энтропиия. Благадаря только этим двум функциям можно достичь хорошего разделения различных видов деятельности. В *Twitter* всего пять основных категорий твитов:

— автоматическая или роботизированная деятельность,
— распространение новостей,
— реклама и продвижение,
— общественные кампании,
— паразитическая реклама.

Третий раздел «Энтропийные меры в случайных графах» посвящён исследованию применения энтропийных мер в случайных графах. Для получения результатов анализа энтропийных мер в случайных графах было проанализировано как изменяется степень энтропии при различных значениях параметра p. Для этого в рамках данного раздела было сгенерировано

множество графов с разным количеством узлов для каждой модели. В целом значение энтропии сети модели Эрдёша-Реньи и Барабаши-Альберт показывают U-образную форму при возрастании параметров N и p. В приложении А приведен разработанный программный код на языке *Python3*.

Четвёртый раздел «Энтропийные меры в финансовых системах» посвящён исследованию применения энтропийных мер в финансовых системах. Были проанализированы результаты для энтропии Шеннона, применена оценка плотности, полученная с помощью метода MAP. Представлены изображения моделей: первая – в разбросе (показатель энтропии и кризиса, как фактический, так и прогнозируемый), а вторая – во временной перспективе (показатель времени и кризиса, как фактический, так и прогнозируемый), а также подчеркивается использование такой модели с точки зрения системы раннего предупреждения.

Энтропия может рассматриваться как мера беспорядка или случайности. Фактически, индекс энтропии, применяемый к вероятности, является максимальным, тогда, когда лежащее в основе распределение вероятности является равномерным. Другими словами, если все элементы векторов $p = (p_1, p_2, ..., p_K)$ равны, то энтропия, связанная с этим вектором, максимальна. В противном случае, если хотя бы один элемент $p_k = 1$, а все остальные равны нулю,то энтропия минимальна.

Обозначение энтропии Шеннона:

$$E_S = \sum_{i=1}^{n} p_i log(p_i),$$

Обозначение энтропии Реньи:

$$E_R = \frac{1}{1-\alpha} log\left(\sum_{i=1}^{n} p_i^{\alpha}\right),$$

Обозначение энтропии Цаллиса:

$$E_t = \frac{1}{1-\alpha}\left(1 - \sum_{i=1}^{n} p_i^{\alpha}\right).$$

Энтропия Шеннона является частным случаем двух других. В частности, в соответствии со значением α меры в уравнениях энтропии Цаллиса и Реньи присваивают более или менее весовые значения хвостам распределе-

ния.

Для каждого из этих определения энтропии были оценены модели, чтобы выявить влияние параметра α энтропии Реньи и Цаллиса. Если E_t – индекс энтропии (любого типа) для распределения возвратов в моменты времени t и C_t является индикатором кризиса в момент времени t, то указанные модели имеют вид:

$$Pr(C_t = 1|E_t) = G(\beta_0 + \beta_1 E_t)$$

а $G(x)$ – логистическая кумулятивная функция плотности модели:

$$G(x) = \frac{e^x}{e^x + 1}.$$

Пятый раздел «Энтропийные меры в социальных сетях» посвящён исследованию применения энтропийных мер в социальных сетях.

Энтропийная классификация может использоваться для обнаружения спама, выявления тенденций, моделирования пользователей, понимания намерений и обнаружения подозрительной активности в социальных сетях, возможно использовать в маркетинговых целях. Было определено пять категорий ретвитов информации: новости, рекламные объявления, общественные кампании, роботизированная деятельность и паразитическая реклама. Метод классификации, основанный на энтропии, не только позволяет характеризовать активность пользователей, но также помогает понять пользовательский контент и отделить популярный контент от обычного или непопулярного контента.

Сделанные наблюдения относительно динамики ретвитов, можно кратко зафиксировать двумя распределениями: распределением по интервалам между твитами и распределением пользователей. Для начала было в ручную проанализировано распределение по интервалам между твитами. Признак человеческой деятельности – широкое распределение с интервалами времени различной длины, которые все одинаково вероятны.

С помощью мер энтропии можно измерить регулярность или предсказуемость временного интервала твитов. Пусть ΔT обозначается как интервал времени между двумя последовательными ретвитами в действии τ_j с возможными значениями $\{\Delta t_1, \Delta t_2, , \Delta t_i, \ldots, \Delta t_{n_T}\}$. Если есть $n_{\Delta t_i}$ временных интервалов длины Δt_i, то $p\Delta T(\Delta t_i)$ обозначает вероятность временного интервала

Δt_i:

$$p\Delta T(\Delta t_i) = \frac{n_{\Delta t_i}}{\sum_{k=1}^{n_T} n_{\Delta t_k}}$$

Энтропия распределения временных интервалов $H_{\Delta T}$ вычисляется как:

$$H_{\Delta T}(\tau_j) = -\sum_{i=1}^{n_T} p\Delta T(\Delta t_i) log(p\Delta T(\Delta t_i))$$

Автоматическая ретвит-активность имеет меньшую энтропию временного интервала и более предсказуема, чем ретвит человека, который более широко распространён и менее предсказуем.

Помимо временного интервала также измеряется распределение числа раз, когда конкретный пользователь ретвитит какой-либо пост. Интересный контент обычно ретвитится один раз каждым пользователем, участвующим в активности твита, спамоподобные действия определяются тогда, когда отдельный человек или небольшая группа неоднократно ретвитят один и тот же пост.

Кампания по накрутке активности, является успешной, тогда, когда в ней участвует много разных пользователей. Подобные спаму характеристики также наблюдаются в рекламных объявлениях.

Энтропийные меры также можно использовать для измерения широты распространения пользователей. Пусть случайная величина F обозначает отдельного пользователя в действии τ_j с возможными значениями $\{f_1, f_2, \ldots, f_i, \ldots, f_{n_F}\}$. Пусть в действии τ_j будут n_{f_i} ретвиты от пользователя f_i. Если p_F определяет массовую функцию вероятности F, так что $p_F(f_i)$ дает вероятность того, что пользователь f_i сгенерирует ретвит, то:

$$p_F(f_i) = \frac{n_{f_i}}{\sum_{k=1}^{n_F} n_{f_k}}$$

Энтропия пользователя H_F определяется как:

$$H_F(\tau_j) = -\sum_{i=1}^{n_F} p_F(f_i) log(p_F(f_i))$$

Временной интервал $H_{\Delta T}(\boldsymbol{\tau}_j)$ и энтропия пользователей $H_F(\boldsymbol{\tau}_j)$ могут использоваться для классификации ретвит-активности контента в *Twitter*. Эта классификация поможет не только идентифицировать различные динамические действия, происходящие в *Twitter*, но также это представляет из себя ценную информацию о природе происхожения данного контента.

В подразделе «Результаты» демонстрируются результаты классификации информации с помощью энтропии.

Роботизированная деятельность. Два основных типа автоматизированного написание твита – это службы автоматического твита и службы планирования твитов. Есть две категории авто-твита. Первый возникает, когда человек подписывается на автоматическую услугу, которая отправляет сообщения в профиль пользователя от его имени.

Важная или интересная информация. Этот класс состоит в основном из новостей и блогов, а также из нескольких популярных общественных кампаний. Интересная информация характеризуется сопоставимой (обычно высокой) пользовательской и временной энтропией. Это разделение достаточно значимое, оно основанно на популярности.

Реклама и спам. Рекламные объявления и спам отличаются низкой энтропией пользователя и низкой или высокой временной энтропией. Это нежелательные рекламные объявления, которые никогда не ретвитятся ни одним пользователем, кроме автора объявления.

Общественные кампании. Кампании идентифицируются по низкой пользовательской энтропией и очень высокой временной энтропией. В наборе данных, помеченном вручную, очень мало кампаний. Из-за значительного совпадения характеристик кампаний с рекламой или рекламными акциями отличить кампанию от рекламы сложно даже вручную.

Паразитная реклама. Идентифицировать паразитную рекламу точно очень сложно. Одной из возможных причин может быть их паразитическая природа, когда они не имеют собственных характерных особенностей, но принимают характеристики профиля пользователя.

В этом разделе данной ВКР была охарактеризована динамика активности в такой сложной системе как *Twitter*, по энтропии пользователей и энтропии временых интервалов была охарактеризована динамика активности.

Показано, что этих двух функций достаточно, чтобы разделить активность пользователей на различные классы. Данный метод является вычислительно эффективным и масштабируемым, не зависит от содержимого и языка контента.

ЗАКЛЮЧЕНИЕ

В ходе данной выпускной квалификационной работы магистра в разделе 1 выполнен анализ современных публикации по темам «меры энтропии», «модели случайных графов» и «энтропия сети». В разделе 2 изучена область применения энтропийных мер.

В разделе 3 данной работы изучены энтропийные меры в случайных графах. Проанализирован алгоритм моделирования сетей вида *small-world network*. Основываясь на энтропии сети, исследована тенденция энтропии сети в процессе создания случайных графов. Случайные графы являются основой для рассмотрения безмасштабных сетей. Практическая реализация применения энтропии представлена в приложении А на языке *Python3*.

В ходе данной ВКР было исследовано две области применения энтропии в реальных сетях, – экономическая система и социальная система.

При исследовании возможности применения энтропийных мер в экономеческой системе в качестве сети исследования была взята финансовая сеть. Вторая область применения энтропийных мер – социальная. Здесь в качестве исследуемой модели послужила социальная сеть *Twitter*. .

В разделе 4 исследована возможность применения энтропийных мер для предупреждения системных рисков в сложных сетях, а именно в финансовой сети. После применения энтропии Шеннона, выяснилось, что энтропия может использоваться для финансовых моделей, в которой агенты на финансовом рынке реагируют на проблемы. Практическая реализация применения энтропии в финансовой системе на языке *MATLAB* приведена в приложении Б.

В разделе 5 изучается поведение социальной сети *Twitter*. Представлена модель для описания роста безмасштабных сетей. Модель можно применять для определения оптимальной стратегии защиты от вредоносных программ и распространения спама, использовать в области маркетинга. По энтропии пользователей и энтропии временых интервалов была охарактеризована динамика активности. Показано, что этих двух функций достаточно, чтобы

разделить активность пользователей на различные классы. Данный метод является вычислительно эффективным и масштабируемым, не зависит от содержимого и языка контента.

Энтропийная классификация может использоваться для обнаружения спама, выявления тенденций, моделирования пользователей, понимания намерений и обнаружения подозрительной активности в социальных сетях, возможно использовать в маркетинговых целях. Было определено пять категорий ретвитов информации: новости, рекламные объявления, общественные кампании, роботизированная деятельность и паразитическая реклама. Метод классификации, основанный на энтропии, не только позволяет характеризовать активность пользователей, но также помогает понять пользовательский контент и отделить популярный контент от обычного или непопулярного контента. Программная реализация на языке *Python3* представлена в приложении В.

Отдельные части магистерской работы были представлены на конференции:

— 16 апреля 2019 10-я научно-практическая конференция, посвященная 110-летию саратовского государственного университета "Presenting Academic Achievements to the World" тема статьи: "Use of Korner entropy in complex networks". Сертификат об участии приложен в приложении Г.

— 24 апреля 2020 студенческая научная конференция, тема доклада «Модель безмасштабной сети *Twitter*».

Основные источники информации:

1. Erdös P., Rényi A. On the evolution of random graphs // Научный журнал Math. Inst. Hungar. Acad. Sci. — 1960. - Т. 5. - С. 17–61. - Сведения доступны также по Интернет: http://leonidzhukov.net/hse/2014/socialnetworks/papers/erdos-1960-10.pdf (дата обращения: 25.10.2018). - Яз. англ.
2. Bollobás B. Random Graphs // Издательство Кэмбриджского унив. - 2001. - С. 520.
3. Barabási L.-A., Albert R. Emergence of scaling in random networks // Научный журнал Science. - 1999. - Т. 286. - С. 509–512.- Сведения доступны также по Интернет: https://barabasi.com/f/67.pdf (дата обращения:

25.10.2018). - Яз. англ.

4. Barabási L.-A., Albert R., Jeong H. Scale-free characteristics of random networks: the topology of the world-wide web // Научный журнал Physica A. - 2000. - С. 69–77.
5. Albert R., Jeong H., Barabási L.A. Diameter of the world-wide web // Научный журнал Nature. - 1999. - С. 130–131. - Сведения доступны также по Интернет: https://barabasi.com/f/65.pdf (дата обращения: 25.10.2018). - Яз. англ.
6. G. Bianconi, P. Pin, M. Marsili. Assessing the relevance of node features for network structure // Научный журнал PNAS. - 2009. - Т. 106. - №28. - 11433.
7. K. Anand, G. Bianconi. Entropy measures for networks: Toward an information theory of complex topologies // Научный журнал Physical Review E. - 2009. - Т. 80. - 045102.
8. G. Bianconi. Entropy of network ensembles // Научный журнал Physical Review E. - 2009. - Т. 79. - 036114.
9. Körner, J. Coding of an information source having ambiguous alphabet and the entropy of graphs. Prague, Czech Republic – 1973. – С. 411–425.
10. Клод Шеннон. Математическая теория связи // Научный журнал The Bell System Technical Journal. - 1948. - Т. 27(3). - С. 379-423.