ТЕОРЕТИЧЕСКИЕ ПОИСКИ И ПРЕДЛОЖЕНИЯ

УДК 311

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРЕДСКАЗАНИЯ РАСПРОСТРАНЕНИЯ ИНФЕКЦИИ В СЕТИ

П.А. Сулимов
Центральный банк РФ
E-mail: sulpav@yandex.ru

Запуск в 2004 г. Facebook послужил толчком для исследования вопроса, как люди взаимодействуют друг с другом в рамках социальной сети, в которой они состоят. С тех пор прошло уже более 10 лет и появилось множество тематических социальных сетей: Twitter, Instagram, LinkedIn, Flickr и т.д. Во всех перечисленных социальных сетях люди обмениваются какой-либо информацией: фотографиями, ссылками, контактами и пр. Информация – своего рода вирус, передающийся от человека к человеку. Соответственно распространение информации в социальной сети рассматривается автором с точки зрения модели заражения (epidemics in social network). В работе ставится задача предсказания порога эпидемии (пороговой характеристики сети, при превышении которой сеть гарантированно оказывается полностью зараженной) в момент времени t+1 на основании исторических данных за периоды t, t–1 и ранее. Для решения поставленной задачи необходимо знать, как поведет себя сеть в момент времени t+1, будет ли граф сети связным, какие связи разорвутся, а какие появятся и т.д. Ведь именно этим определяются скорость распространения инфекции по сети и порог эпидемии. Соответственно возникает проблема Link Prediction Problem, которая решается методами машинного обучения (Random Forest, Support Vector Machines) путем отнесения пар вершин к классам соединенных и несоединенных и предсказания класса пары вершин в момент времени t+1 на основании топологических и факторных характеристик узлов сети. Таким образом, результатом исследования является алгоритм прогнозирования распространения инфекции в социальной сети при помощи методов машинного обучения.

Ключевые слова: социальная сеть, модель заражения, Link Prediction Problem, Random Forest.

METHODS OF MACHINE LEARNING TO PREDICT THE SPREAD OF THE INFECTION IN THE NETWORK

P.A. Sulimov
Central Bank of the Russian Federation
E-mail: sulpav@yandex.ru

The launch of Facebook in 2004 gave rise to research of the question, how people interact with each other within a social network. Since then more than 10 years passed, and many thematic social networks appeared: Twitter, Instagram, LinkedIn, Flickr etc. People exchange any information (photos, links, contacts etc.) in all listed social networks. Infor-

mation is some kind of virus which is transferred from person to person. Respectively, the author considers distribution of information in a social network from the point of view of model of infection (epidemics in social network). The paper sets the goal of epidemic threshold prediction (threshold characteristic of a network, above which the network is surely completely infected) in the time point t+1 on the basis of historical data for the periods of t, t–1 and earlier. For the solution of the set goal it is necessary to know how the network will behave in the t+1 time point, whether the network graph is connected, what links will be broken and what will appear etc. All of the above define the speed of spread of an infection in networks and epidemic threshold. Respectively, the Link Prediction Problem emerges, which is solved by methods of machine training (Random Forest, Support Vector Machines) by referring pairs of nodes to classes connected and not connected, and predictions of class of pair of nodes in the t+1 time point on the basis of topological and factorial characteristics of knots of network. Thus, the algorithm of forecasting of spread of infection in a social network by means of methods of machine training is the result of the research.

Keywords: the social network model of infection, Link Prediction Problem, Random Forest.

ВВЕДЕНИЕ

Запуск в 2004 г. Facebook послужил толчком для исследования вопроса, как люди взаимодействуют друг с другом в рамках социальной сети, в которой они состоят. С тех пор прошло уже более 10 лет, и появилось множество тематических социальных сетей: Twitter, Instagram, LinkedIn, Flickr и т.д. Во всех перечисленных социальных сетях люди обмениваются какой-либо информацией: фотографиями, ссылками, контактами и пр.

Информация – своего рода вирус, передающийся от человека к человеку. А значит, чтобы понять, как происходит распространение информации в сети, можно рассмотреть данный вопрос с точки зрения модели заражения (epidemics in social network). Однако модели распространения инфекции в сети в основном рассматриваются с предположением о неизменной численности системы. Но даже при условии появления динамики (новых узлов сети) отсутствует четкий механизм образования связей между участниками сети. А ведь в реальной жизни, чтобы понять, с какой вероятностью передастся вирус от человека к человеку, необходимо прежде всего понимать, а существует ли вообще контакт между этими людьми.

Соответственно для правильного прогнозирования распространения инфекции в сети следует решить проблему предсказания связей.

В первом разделе подробно рассматривается модель заражения, наиболее полно отражающая факт распространения по сети информационного импульса и продуктовых новинок от одного агента к другому (механизм последовательного инфицирования соседей). Во втором разделе исследуется проблема предсказания связей в социальной сети. Приводится формальная постановка проблемы, описываются существующие техники ее решения. В заключительном, третьем разделе, проводится исследование на примере сети Flickr.

С целью проведения анализа в качестве математического аппарата в работе используется теория графов и метод машинного обучения Random Forest. Код программы, делающей вычисления, пишется в среде R.

1. МОДЕЛЬ ЭПИДЕМИИ SI

Для того чтобы представить, как устроен процесс заражения сети, рассмотрим одну из самых популярных и применимых на практике моделей – SI (suspectible – infected) [4].

Опишем предпосылки модели SI:

– здоровое население, чувствительное к заражению, $S(t)$;

– зараженное население, $I(t)$;

– численность населения постоянна, $S(t) + I(t) = N$;

– отсутствие выздоровления;

– моментальное заболевание человека в случае инфицирования;

– нормальное распределение людей по системе;

– обозначим интенсивность заражения через λS;

– показатель инфицирования пропорционален числу зараженных, т.е. $\lambda = \beta I$.

Пара обыкновенных дифференциальных уравнений, описывающих данную модель, имеет вид:

$$\frac{dS}{dt} = -\beta I(t)S(t);$$

$$\frac{dI}{dt} = \beta I(t)S(t).$$

Однако поскольку $N = S(t) + I(t)$, то эта пара уравнений эквивалентна следующей:

$$\frac{dS}{dt} = -\beta I(t)S(t);$$

$$\frac{dI}{dt} = \beta I(t)(N - I(t)).$$

Данное дифференциальное уравнение является логистическим уравнением роста.

Разделив обе части уравнения на $I(t)(N - I(t))$, получим

$$\frac{1}{I(t)(N - I(t))}\frac{dI}{dt} = \beta.$$

Проинтегрируем полученное уравнение:

$$\int_0^t \frac{1}{I(t)(N - I(t))}\frac{dI}{dt}d\tau = \int_0^t \beta d\tau;$$

$$\int_{I(0)}^{I(t)} \frac{1}{u(N - I(t))}du = \int_0^t \beta d\tau;$$

$$\frac{1}{N}\int_{I(0)}^{I(t)} \frac{1}{u} + \frac{1}{N - u}du = \int_0^t \beta d\tau;$$

$$[\ln(u) - \ln(N - U)]_{u=I(0)}^{I(t)} = \beta d\tau.$$

После некоторых математических преобразований получаем итоговое уравнение

$$I(t) = \frac{I(0)N}{I(0) + (N - I(0))e^{-\beta Nt}}.$$

Данное уравнение описывает логистическую кривую. Основной вывод модели заключается в том, что при t, стремящемся к бесконечности, I стремится к значению N, т.е. в перспективе инфицированным окажется каждый.

Чаще всего исследователи рассматривают модели SI с предпосылкой о неизменной численности системы. При появлении предпосылки об изменении состава популяции обычно вводятся постоянные коэффициенты выбытия и рождаемости (появления новых членов). Но даже при условии появления динамики связи между «старыми» участниками сети остаются неизменными (т.е. новые связи между ними не появляются).

По мнению автора, отсутствие динамики связей между агентами в моделях SI является серьезным недостатком. И для правильного прогнозирования распространения инфекции в сети следует, прежде всего, решить проблему предсказания связей.

2. ПРОБЛЕМА ПРЕДСКАЗАНИЯ СВЯЗЕЙ (LINK PREDICTION PROBLEM)

Когда говорится о социальной сети, имеется в виду динамичная развивающаяся система, или же, выражаясь математическим языком, граф с изменяющимися во времени узлами и ребрами. И вправду, невозможно себе представить, чтобы количество участников социальной сети и связей между ними оставалось постоянным – кто-то покидает сеть, кто-то приходит; некоторые люди теряют свои контакты и в то же время находят новые.

И если предсказать появление новых участников в социальной сети не представляется возможным, то попытки «угадать», между какими узлами в ближайшем будущем возникнет связь, делаются уже довольно давно.

Описанная выше процедура получила в англоязычной литературе название Link Prediction (дословный перевод – предсказание связей). Изучение данной проблемы было вызвано прежде всего тем, что любая существующая социальная сеть является «неполной» в том смысле, что чаще всего исследователям доступна только часть информации об участниках сети и связях между ними, которую возможно извлечь с социальных платформ. Более того, неполнота заключается также в том, что внутри сети есть участники, которые, возможно, знают друг друга, но в то же время не обозначены как «друзья» внутри сети.

Таким образом, решение проблемы предсказания связей в сети способно принести пользу сразу в нескольких областях исследования социальных сетей. С одной стороны, результаты предсказаний могут носить рекомендательный характер, если речь идет о друзьях в Facebook, потенциальном сотрудничестве бизнес-партнеров, покупках в интернет-магазинах и т.д. С другой стороны, предсказанные возможные связи могут служить сигналом для действия правоохранительных органов (если речь идет о связях в криминальных структурах).

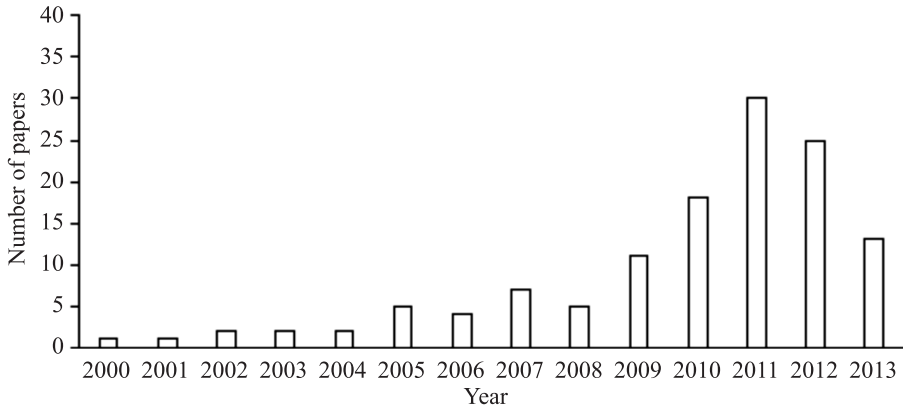

Рис. 1. Количество опубликованных работ с ключевыми словами «Link Prediction» в период 2000–2013 гг.

О важности решения проблемы предсказания связей в сети можно судить и по тому, сколько работ на данную тему было опубликовано в период 2000–2013 гг. (рис. 1).

Несмотря на большое количество публикаций, реальным толчком для исследований в области предсказания связей стала работа Liben-Nowell & Kleinberg [3], опубликованная в 2003 г., в которой авторы описали методику прогнозирования связей, основанную на топологических особенностях социальной сети. Пару лет спустя вышла работа Hasan & Zaki [1], посвященная применению методов машинного обучения в LPP (Link Prediction Problem). Для сравнения были взяты популярные методы классификации (деревья решений, SVM – машина опорных векторов, метод ближнего соседа, наивный Байес и т.д.). Очевидно, что список работ по исследованию LPP не ограничивается двумя представленными трудами, однако эти две работы являются фундаментом для описания существующих способов предсказания связей в сети.

Далее приведена формальная постановка проблемы LPP, описаны самые популярные методики, используемые для предсказания связей, а также представлена модификация существующих моделей прогнозирования связей в сети применительно к анализу временных рядов графов.

2.1. Постановка проблемы предсказания связей

Пусть существует социальная сеть, задаваемая графом $G(V, E)$ в определенный момент времени t. Тогда задача предсказания связей сводится к тому, чтобы предсказать связи, которые возникнут в дискретный момент времени $t'(t' > t)$, или же выявить ненаблюдаемые связи в текущий момент времени. К примеру, у нас есть сеть из 5 друзей (рис. 2).

В момент времени t Алена знает Диму, Лену, Мишу и Максима. Однако в момент времени t' Алена может познакомить между собой Мишу и Максима, или же Максима и Диму (эти связи обозначены пунктиром). Или же может оказаться так, что Максим и Миша уже знакомы, но просто мы об этом не знаем (ненаблюдаемая связь).

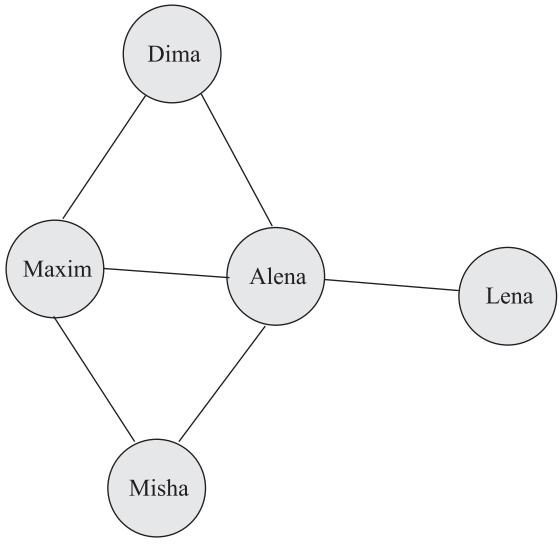

Рис. 2. Мини-социальная сеть из 5 друзей

Для того чтобы понять, насколько велика вероятность того, что между двумя выбранными вершинами образуется связь, были разработаны два основных метода: подход, основанный на схожести вершин (*similarity – based approach*), и подход, основанный на обучении (*learning – based approach*).

Similarity – based approach основан на том, что исследователем высчитываются коэффициенты «похожести» для каждой пары несоединенных вершин (о разнообразии коэффициентов «похожести» и методиках их расчета будет подробно рассказано далее). Затем, посчитанные величины ранжируются в порядке убывания (так как большее значение коэффициента означает бóльшую вероятность для двух вершин быть соединенными). В итоге те пары вершин, чьи коэффициенты оказались в верхушке списка, имеют наибольшую вероятность соединения.

Learning – based approach исходит из того, что задача предсказания связей в сети определяется как задача бинарной классификации (т.е. 0 – если вершины не соединены, 1 – иначе). Тогда для решения проблемы LPP используется какой-либо метод машинного обучения (или же несколько методов). В уже упомянутой работе Hasan & Zaki [1] были проведены исследования на двух сетях: BIOBASE (данные о научных публикациях в сфере биологии) и DBLP (исследования в сфере компьютерных наук). В результате наибольшую точность в предсказании связей на данных BIOBASE показали деревья решений и бэггинг, а на данных DBLP лучший результат показал SVM.

Для применения вышеозначенных методов машинного обучения к задачам LPP берется «срез» сети за период времени $[t, t']$ в качестве тренировочной выборки, и сеть в период $[t'', t''']$, $t'' > t'$ в качестве тестовой выборки. Результатом обучения модели является классификатор, определяющий вероятность возникновения связи между вершинами.

Learning – based approach является более практичным подходом в сравнении с *similarity – based approach*, так как последний только ранжирует посчитанные коэффициенты и определение того, в каком месте провести линию, выше которой находятся пары вершин, соединение которых наиболее вероятно, остается на усмотрение исследователя. Таким образом, результат становится слишком субъективным. Поэтому в дальнейшем исследовании будет изучаться проблема LPP только через призму *learning – based approach*.

2.2. Метрики в моделях предсказания связей

Существует три основные категории метрик, которые используются для прогнозирования связей в сети. Опишем основные метрики из каждой категории.

Показатели схожести. Из всех метрик, данная метрика – единственная, не использующая особенности сети, а лишь свойства самих участников сети.

К примеру, в уже упомянутой работе Liben-Nowell & Kleinberg [3] авторы проводили исследования на сети соавторов научных работ из разных областей. Были изучены работы из пяти различных областей физики и в каждой из них в качестве вершин были авторы работ, а в качестве ребра между вершинами – написанная вместе статья (предполагалось, что между двумя вершинами существует ребро, если авторы написали вместе хотя бы одну статью). Из графа сети соавторов для исследования был выбран подграф как в тренировочном периоде (брались авторы, которые в период 1994–1996 гг. написали хотя бы три статьи), так и в тестовом периоде (аналогично брались авторы в период 1997–1999 гг.).

В данном случае в качестве одной из метрик можно было выбрать «совпадение ключевых слов в написанных двумя авторами статьях». Эта метрика являлась бы показателем схожести, так как она не учитывает особенности топологии или свойств графа, однако использует характеристики непосредственно самих авторов (несмотря на то, что в описанной статье эта метрика не применялась, она стала использоваться в аналогичных работах по исследованию сетей соавторов, написанных позже).

Топологические особенности. Liben-Nowell & Kleinberg также использовали в своем исследовании [3] многие популярные топологические метрики и сравнили их «предсказательную силу». Ниже приведем использованные в статье метрики, а также некоторые другие.

Метрики соседства. Одна из особенностей социальных сетей заключается в том, что люди склонны заводить контакты с теми, у кого с ними уже есть общие знакомые. На этом свойстве социальных сетей и построены следующие метрики.

– Общие соседи (*CN*). Если в сети есть вершины x и y, то для них можно определить метрику общего соседства:

$$CN(x, y) = \left|\Gamma(x) \cap \Gamma(y)\right|,$$

где $\Gamma(x)$ – число соседей y вершины x (т.е. число инцидентных вершин), а $\Gamma(y)$ – соответственно число соседей для вершины y. Данная метрика вполне логична с бытовой точки зрения – чем больше у двух незнакомых людей общих знакомых, тем более вероятно то, что эти два человека также знают и друг друга. Однако проблема этого показателя в том, что он не нормирован, т.е. сравнивать его с аналогичными, но рассчитанными для других вершин нельзя. Эту проблему сравнения решает следующая метрика.

– Коэффициент Жаккарда (*JC*)

$$JC(x, y) = \frac{\left|\Gamma(x) \cap \Gamma(y)\right|}{\left|\Gamma(x) \cup \Gamma(y)\right|}.$$

В данном случае предыдущий показатель числа общих соседей нормируется на суммарное число соседей у двух людей (и теперь данный коэффициент можно сравнивать с другими, т.е. можно делать вывод о том, что если $JC(x_1, y_1) > JC(x_2, y_2)$, то вероятность соединения первой пары вершин больше, чем для второй).

– Адамик/Адар (AA). Идея данного коэффициента немного отличается от двух предыдущих, так как в данном случае коэффициент представляет собой следующую сумму:

$$AA = \sum_{g \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log \Gamma(g)}.$$

Важным является то, сколько связей у общих знакомых двух друзей.

Метрики расстояния. Кроме общих соседей среди топологических особенностей можно выделить расстояние между двумя вершинами.

– Катц. Очевидно, что чем короче расстояние от одной вершины до другой, тем более вероятно то, что рано или поздно они свяжутся. А если путей из вершины x в вершину y будет несколько, то это только увеличит вероятность возникновения связи. Однако наличие длинного пути между вершинами вносит меньший вклад в вероятность соединения, чем наличие более короткого пути. Коэффициент Катца учитывает все вышеперечисленные особенности:

$$\text{Katz}(x, y) = \sum_{l=1}^{\infty} \beta^l \left| path_{(x,y)}^{(l)} \right|,$$

где $path_{(x,y)}^{(l)}$ – путь длины l между вершинами x и y, $\beta > 0$ – штраф за длину пути (чем длиннее путь, тем меньший вклад в суммарный коэффициент он будет вносить).

Случайное блуждание.

– Время достижения. Данный коэффициент показывает, какое количество шагов требуется, чтобы при помощи случайного блуждания добраться из вершины x в вершину y:

$$HT(x, y) = 1 + \sum_{w \in \Gamma(x)} P_{x,w} HT(w, y).$$

– Время сообщения. Предыдущий показатель учитывает только количество шагов из вершины x в вершину y. Однако при случайном блуждании это количество не равно количеству шагов из вершины y в вершину x:

$$CT(x, y) = HT(x, y) + HT(y, x).$$

– SimRank. Для следующей метрики делается предпосылка о том, что две вершины наиболее вероятно будут соединены в том случае, если они соединены с похожими вершинами:

$$\text{simRank}(x, y) = \begin{cases} 1, x = y \\ \gamma \cdot \dfrac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} \text{simRank}(a, b)}{\left|\Gamma(x)\right| \left|\Gamma(y)\right|}, \text{ иначе} \end{cases}$$

где $\gamma > 0$ – штраф за расстояние. Таким образом, можно сказать, что данная метрика показывает, насколько быстро встретятся две точки, если будут

«случайно блуждать» по графу – при этом первая точка начинает блуждание из вершины x, а вторая – из вершины y соответственно.

Метрики, основанные на свойствах вершин.

– Предпочтительное присоединение. Идея данной метрики основана на том, что вершины в сети стремятся присоединиться к тем вершинам, у которых уже есть большое количество связей

$$PA(x, y) = |\Gamma(x)||\Gamma(y)|.$$

– Показатель кластеризации

$$CC(x, y) = CC(x) + CC(y).$$

Смешанный социальный показатель. Идея данного показателя схожа с идеей коэффициента Адамика/Адара о том, что общие соседи вносят различный вклад в вероятность соединения двух узлов сети:

$$LCW(x, y) = \sum_{z \in \Gamma(x) \cup \Gamma(y)} w(z)^{\beta},$$

где $w(z)$ – показатель центральности вершины z, $\beta > 1$ – штраф за низкое значение центральности вершины.

На этом список метрик не заканчивается, однако приведенных выше показателей достаточно для того, чтобы сформировать полную картину о возможных метриках для предсказания связей между узлами в сети.

Как уже было сказано, в статье [3] авторы сравнивали «предсказательную силу» различных метрик. Основные результаты сравнения представлены на рис. 3.

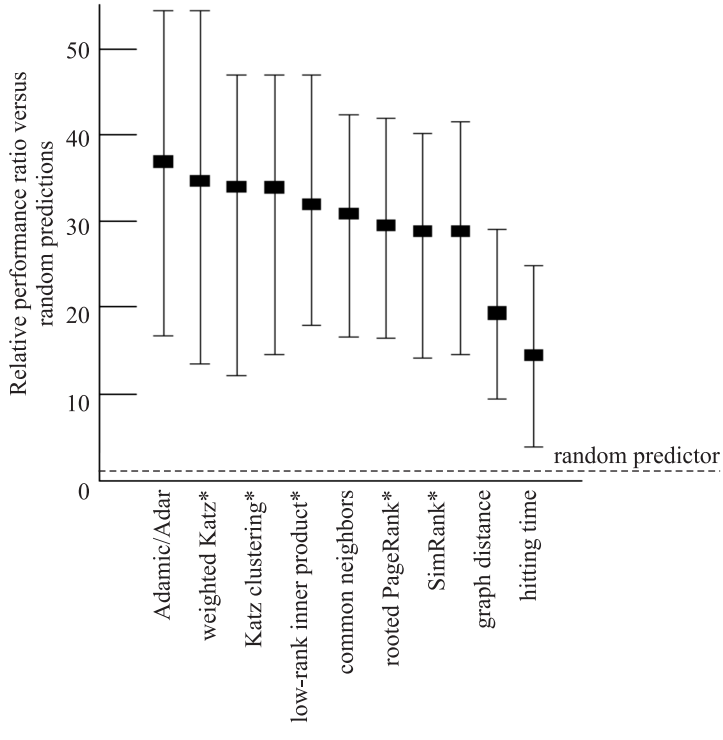

Рис. 3. Сравнение предсказательной силы метрик в исследовании *Liben-Nowell & Kleinberg*

Несмотря на представленное ранжирование метрик по «предсказательной силе», на практике использование той или иной метрики зависит от решаемой задачи, и чаще всего исследователями используются несколько метрик сразу, или же их комбинации. Однако стоит быть осторожным при комбинировании метрик, так как совмещение в одной модели метрик, схожих по методологии расчета, может привести к смещенной оценке.

2.3. Динамические модели предсказания связей

В большинстве современных исследований авторы рассматривают вариант, когда изучаемая сеть представлена в виде неориентированного не взвешенного графа, и при этом чаще всего сеть является гомогенной (т.е. все вершины равноправны между собой). Несмотря на то, что реальные сети нередко бывают гетерогенными (это означает, что сеть является мультиагентской: к примеру, сеть P2P-Lending с вершинами – кредиторами и вершинами – заемщиками), данные сети сложны для изучения, и в данной работе рассматриваться не будут.

Особый интерес вызывает вопрос, какие подходы используются для прогнозирования связей в сети в момент времени $t + 1$ при наличии ретроспективной информации на момент времени t. Дело в том, что, как было описано выше, авторы многих статей используют подход *snapshot*: берется «срез» сети за период времени $[t, t']$ в качестве тренировочной выборки, и сеть в период $[t'', t''']$, $t'' > t'$ в качестве тестовой выборки (если говорить о машинном обучении). Таким образом, составляется предиктор, способный по входящей информации предсказывать связи в сети.

Проблема такого подхода заключается в том, что сеть не рассматривается как динамическая система. Это означает, что сеть за некий промежуток времени интерпретируется как пространственный объект. Хотя если взять несколько «срезов» по времени и построить модель зависимости во времени, то мы получим желаемую интерпретацию предсказания связей в форме *temporal link prediction*.

Изучению проблемы предсказания связей, используя теорию временных рядов, посвящено довольно немало статей. К примеру, в статье [2] авторами Huang & Lin берется T штук «срезов» одной и той же сети в разные моменты времени, задаваемые графами $(G_1, \ldots, G_T)$. Для простоты полагается, что со временем в сети не появляются новые узлы (те, которые появляются, просто исключаются из выборки) и их количество равно N. Тогда можно задать для каждого фиксированного момента времени t симметричную матрицу смежности с элементами $M_t(i, j)$, отвечающими числу взаимодействий между вершинами i и j за единицу времени t.

В качестве данных для исследования был выбран Enron E-mail Data Set. В выборке представлено 252 759 писем работников компании Enron, которые писали друг другу 151 человек в течение трех лет (за единицу времени был выбран месяц). Был предложен смешанный подход к прогнозированию связей в сети: с одной стороны, для каждого месяца авторами были посчитаны по пространственным данным (т.е. внутри одного месяца) различные метрики типа описанных выше (общих соседей, Адамик/Адар, Катц, Жаккард). С другой стороны, авторы предположили, что наличие связи в

момент времени t между узлами i и j можно описать при помощи стандартной ARIMA–модели, используя данные из матриц смежности (т.е. наличие связи между вершинами может объясняться интегрированной моделью авторегрессии – скользящего среднего).

Алгоритм предсказания связей в общем виде выглядит следующим образом:

1. Преобразовать ряд матриц смежностей в одну, отвечающую статичному графу: составить матрицу $M_{1\sim T} : M_{1\sim T}(i, j) = \sum_{t=1}^{T} M_t(i, j)$, после этого составить матрицу $M^*_{1\sim T} : M^*_{1\sim T}(i, j)$, если $M_{1\sim T}(i, j) > 1$ (иначе $M^*_{1\sim T}(i, j) = 0$).

2. Используя одну из описанных выше метрик, сделать предсказание на период $T + 1$: $S_S = g(M^*_{1\sim T})$, где S_S – матрица вероятностей образования связи между узлами i и j.

3. Создать ряд частоты взаимодействий между узлами i и j $\{x_{ijt}\}$, где $x_{ijt} = M_t(i, j)$.

4. Перебрать все возможные варианты модели ARIMA(p, d, q) для $p = 0, 1, 2, 3$, $d = 0,1$, $q = 0, 1, 2, 3$ и спрогнозировать значение для x_{ijT+1}.

5. Выбрать лучшую модель по критерию Акаике.

6. $S_T(i, j) = \Pr(\hat{x}_{ijT+1} > 1)$.

7. Провести нормализацию: $S_S = \dfrac{S_S}{\text{sum}(S_S)}$, $S_T = \dfrac{S_T}{\text{sum}(S_T)}$.

8. Сделать предсказание связей на основе комбинированного метода статической и динамической моделей прогнозирования связей:

$$S(i, j) = \left(S_S + \frac{\min(S_S)}{\alpha} \right) \cdot \left(S_T + \frac{\min(S_T)}{\alpha} \right), \ \alpha > 1.$$

По итогам исследования комбинированная показала результаты лучше, чем в отдельности статическая и динамическая модели. А комбинация TS (метод time-series – посчитано из ARIMA–модели) и KZ (коэффициент Катца) дала наименьшую ошибку предсказания связей в сети.

Авторы Ahmed & Chen статьи [5] предлагают другую модификацию стандартного подхода к прогнозированию связей в сети (статической модели):

1. Составить модифицированную матрицу смежностей

$$M^*_t = \begin{cases} 1, \text{если } i = j \\ \sum_{t=1}^{T} \delta^{T-(t-1)} M_t(i, j) \end{cases}, \text{ иначе}$$

где $0 < \delta < 1$ – демпингующий фактор, придающий больше значимости более поздней информации.

2. Используя одну из описанных выше метрик, сделать предсказание на период T +1: $S_t = g(M^*_{1\sim T})$, где S_t – матрица вероятностей образования связи между узлами i и j.

3. Для $t = 1, \ldots, T$ построить модель предсказания связей на основе данных в форме временных рядов:

$$S_T = S_T + \delta^{T-(t-1)} S_t.$$

Как можно заметить, основное отличие от предыдущей модели заключается во введении демпингующего фактора. По результатам исследования на тех же данных Enron лучшую прогностическую точность также показала комбинация динамической модели и статической оценки коэффициента Катца.

Далее в работе автором будет представлен свой метод прогнозирования связей в сети с использованием динамических моделей.

3. ПРЕДСКАЗАНИЕ ПОЯВЛЕНИЯ СВЯЗЕЙ И РАСПРОСТРАНЕНИЯ ИНФЕКЦИИ В СЕТИ FLICKR

Для целей прогнозирования связей в сети можно выбрать любую социальную сеть, так как для предсказательного моделирования требуется лишь наличие динамики (появления новых вершин и связей со временем) и взаимодействия между вершинами (в качестве взаимодействия рассматривается установление контакта между двумя пользователями).

В проводимом исследовании будут использоваться данные сети Flickr.

3.1. Построение временного ряда графов сети Flickr

Flickr – социальная сеть, представляющая собой фото- , а также видеохостинг. Сеть была создана в 2004 г. компанией Ludicorp. На данный момент сервис принадлежит компании Yahoo!. Ключевое отличие данной социальной сети от самого популярного приложения по обмену фотографий на мобильных устройствах Instargram (запущен в 2010 г.) заключается в том, что во Flickr пользователи размещают полноформатные фотографии, а не квадратные уменьшенные фото (часто с применением фильтров), как в Instargram.

Идеей сети Flickr является объединение профессиональных фотографов во всем мире (т.е. сети присуще самое важное для нас свойство – возможность установления контактов между пользователями), которые делятся своими фотографиями, делая под ними подписи (возможно, даже истории о том, где и как была сделана фотография), а также отмечая на них людей и ставя маркеры – так называемые в современном интернет-сообществе хэштэги (обозначается #). Именно хэштэги позволяют участникам сети находить интересующие их фотографии (например, если пользователь интересуется фотографиями животных и, в частности, фотографиями сов, то ему достаточно вбить в поисковой строке сети Flickr ключевые слова «животные» и «сова», чтобы найти нужные фотографии).

В качестве входных данных для исследования были взяты два списка. В одном из них находятся данные о пользователях сети Flickr – идентификационный номер пользователя, время регистрации в сети (в секундах), город и страна проживания и т.д. В другом – информация о связях, установленных между пользователями (идентификационные номера обоих пользователей и время установления связи).

Для изучения, как и в модели, описанной в предыдущей главе, возьмем T штук «срезов» сети в разные моменты времени, задаваемые графами $(G_1, \ldots, G_T)$. На основе этой информации нам необходимо сделать прогноз связей в момент времени $T + 1$.

Прежде всего следует выделить T графов из сплошного временного ряда связей, имеющегося во входных данных. Стоит отметить, что пользователи начали активно регистрироваться в сети лишь с 26743615-й секунды, и именно эту точку мы берем за начало отсчета для построения графов. В нашем распоряжении имеются данные вплоть до 81432010-й секунды (т.е. данные о динамике сети почти за 2 года). Однако опытным путем можно убедиться в том, что машинные возможности для анализа графов не бесконечны (при создании таблицы для графов в момент времени больше, чем 30000000 секунд, возникает нехватка памяти, так как требуется разместить вектор объема свыше 1,54 Гб (все расчеты производятся в среде *R* (http://r-project.org)).

Таким образом, нами рассматривается промежуток начального роста сети Flickr. В силу того, что для обработки доступны лишь графы до 30000000-й секунды, нами задается цикл, который строит графы от 26743615-й до 28903615-й секунды с шагом в 86400 с (т.е. в один день). В итоге мы получаем 25 графов, т.е. в нашей модели $T = 25$. Обозначим граф, созданный для сети к 26743615-й секунде G_1, и так далее вплоть до графа G_{25} для сети к 28903615-й секунде.

Алгоритм построения графов основывается на том, что изначально строятся 25 пустых графов, а затем при помощи цикла в граф добавляются все вершины, зарегистрированные в сети на момент времени t, после чего в граф добавляются все связи, которые установлены также на момент времени t. Данный алгоритм позволяет получить на выходе 25 графов (отметим, что каждый граф является неориентированным, а структура исследуемой сети – гомогенная, т.е. все вершины равноправны между собой).

Имея построенные графы, перейдем к извлечению различных метрик и показателей схожести вершин, на основе которых можно будет построить модель классификации вершин графов (т.е. модель предсказания появления связей в сети в момент времени $T + 1$ на основе имеющейся информации в момент времени T).

В качестве метрик выберем коэффициент Жаккарда (метод соседства), произведение локальных коэффициентов кластеризации (метрика, основанная на свойствах вершин) и значение кратчайшего расстояния между вершинами, взятое со знаком минус (так как наша модель строится по принципу того, что чем больше значение предиктора, тем выше вероятность образования связи). Показатели схожести разнятся в зависимости от задачи, но в целом идея заключается в том, чтобы для числовых характеристик похожести ввести метрику $score(i, j) = |x_i - x_j|$, где x_k – характеристика вершины k (к примеру, если речь идет о сетях P2P – Lending, то в них заемщики устанавливают максимальную ставку процента, под которую они готовы взять кредит, а кредиторы – минимальную, под которую они готовы дать кредит, и вероятность связи между кредитором и заемщиком тем выше, чем меньше модуль разности заявленных одним и другим ставок процента). Для сети Flickr можно было бы использовать в качестве $score(i, j)$ города и/или страны, в которых живут пользователи, однако в силу того, что данные по городам и странам очень зашумлены (много пропусков, а также различные варианты написания названия городов), прогноз по предиктору $score(i, j)$ проводиться не будет.

Предположим зависимость вероятности соединения пары вершин в следующий момент времени не только от значения метрик в предыдущий момент времени, но и в моменты с некоторым временным лагом [6]. Возьмем лаг в 2 периода. Зависимость от метрик с временным лагом объясняется тем, что если наблюдается рост показателя того или иного классификатора со временем, то значит на каждом следующем шаге увеличивается вероятность двух вершин быть соединенными. В то же время помимо разности показателей в периоды $(t-1)$ и $(t-2)$ включаем в модель значение показателя метрики в момент времени t, так как разность показателя во времени может оказаться положительной, но значение самого предиктора будет все еще мало.

Для решения задачи классификации будет использоваться метод машинного обучения Random Forest. И хотя Random Forest – довольно мощный классификатор, он не способен строить модели по данным в форме временных рядов. Соответственно придется преобразовать исходные панельные данные в пространственную структуру. Для каждого построенного графа для каждой пары вершин i и j вычисляются следующие характеристики (табл. 1):

$JC_{t-1}(i,j)$ – показатель метрики Жаккарда для ребра (i,j) в момент времени $t-1$,

$SP_{t-1}(i,j)$ – показатель метрики кратчайшего пути для ребра (i,j) в момент времени $t-1$,

$CC_{t-1}(i,j)$ – показатель метрики произведения локальных коэффициентов кластеризации для ребра (i,j) в момент времени $t-1$,

$\text{Link}_t(i,j)$ – бинарная переменная (наличие связи между вершинами i и j в момент времени t.

Таблица 1

Фрагмент таблицы предикторов и связей, составленной для графа G_1

Edges	Link_1	SP_1	JC_1	CC_1
1_2	0	–2	0,28	0,67
1_3	–1	1	0,14	0,73
1_4	0	–2	0,053	0,3
1_5	0	–3	0	0,33
1_6	0	–Inf	0	0,33

Таким образом, составив подобные таблицы для каждого из 25 графов, нам удается перейти от панельных данных к пространственным.

Теперь специфицируем нашу модель классификатора с полным набором заявленных ранее предикторов:

$$\begin{aligned}\text{Link}_t(i,j) \sim JC_{t-1}(i,j) + [JC_{t-1}(i,j) - JC_{t-2}(i,j)] + \\ + SP_{t-1}(i,j) + [SP_{t-1}(i,j) - SP_{t-2}(i,j)] + \\ + CC_{t-1}(i,j) + [CC_{t-1}(i,j) - CC_{t-2}(i,j)],\end{aligned}$$

где знак «+» не имеет смысла знака сложения, а используется для перечисления независимых переменных, используемых для предсказания наличия связи между вершинами.

Очевидно, что для реализации такой модели при помощи алгоритма классификации Random Forest, необходимо составить для каждого периода времени табл. 2, из которой машина могла бы считывать данные.

Таблица 2

Фрагмент таблицы показателей для прогнозирования $\mathbf{Link_3}(i, j)$

Edges	$Link_3$	SP_2	JC_2	CC_2	$SP_2 - SP_1$	$JC_2 - JC_1$	$CC_2 - CC_1$
1_2	0	–2	0,28	0,67	0	0	0
1_3	1	–1	0,14	0,73	0	0	0
1_4	0	–2	0,047	0,35	0	–0,0023	–0,002
1_5	0	–3	0	0,33	0	0	0
1_6	0	–Inf	0	0,33	0	0	0

Заметим, что для составления табл. 2 прогнозирования $Link_t(i, j)$ берутся графы G_t, G_{t-1} и G_{t-2}, причем из двух первых графов удаляются те вершины, которых нет в последнем.

Таким образом получаем 23 таблицы, которые «склеиваем» по рядам в одну большую таблицу (итоговое число рядов – 74542 штуки). И именно эту финальную таблицу мы будем использовать в качестве входных данных при построении модели классификатором Random Forest.

3.2. Прогнозирование методом Random Forest

Для начала интерпретируем переменную Link как факторную (это необходимо для того, чтобы *R* решал задачу классификации, а не задачу регрессии).

Прежде чем разбивать совокупность на обучающую и тестовую выборки, следует решить проблему несбалансированности классов. Дело в том, что в нашем случае класс вершин, которые не соединены (Link = 0), намного превосходит класс тех вершин, которые соединены (Link = 1), из-за этого возникает проблема несбалансированности классов, т.е. классификатор может определять любую выбранную вершину в класс несоединенных и с большой вероятностью окажется прав.

Чтобы решить эту проблему, сделаем выборку, в которой искусственно занизим количество наблюдений с Link = 0 и одновременно сделаем больше наблюдений с Link = 1. Таким образом, примерно уравняем между собой представителей классов соединенных и несоединенных вершин. После этого разобьем полученную новую совокупность случайно на две выборки – одна обучающая (70 %), другая – тестовая (30 %).

Построим модель на обучающей выборке (будем обозначать количество решающих деревьев через *ntree*, количество предикторов, отбираемых на каждом шаге – через *mtry*) (табл. 3).

Из матрицы ошибок, допущенных при классификации, получаем, что из 2406 несоединенных пар вершин классификатор правильно выявил 2400, а 6 пар несоединенных вершин отнес к соединенным, т.е. ошибка прогноза для несоединенных пар вершин составила $\frac{2400}{2406} \approx 0,25\,\%$. Из 1599 соединенных

Таблица 3

Результат классификации Random Forest на обучающей выборке (*ntree* = 500, *mtry* = 2)

Оценка ошибки OOB			1,1 %
Матрица ошибок			
	0	1	Ошибка классификации
0	2400	6	0,25 %
1	38	1561	2,38 %

Примечание. Здесь и далее в матрице ошибок по строкам – фактические категории, по столбцам – спрогнозированные категории).

пар вершин классификатор правильно выявил 1561, а 38 пар соединенных вершин отнес к несоединенным (см. табл. 3). Таким образом, ошибка для соединенных пар вершин равна $\frac{1561}{1599} \approx 2{,}38\,\%$.

Теперь построим модель на тестовой выборке (табл. 4).

Таблица 4

Результат классификации Random Forest на тестовой выборке (*ntree* = 500, *mtry* = 2)

Оценка ошибки OOB			1,26 %
Матрица ошибок			
	0	1	Ошибка классификации
0	1040	6	0,57 %
1	16	686	2,27 %

Для переобучения моделей характерно то, что ошибка на тестовой выборке превышает ошибку на обучающей выборке. В нашем случае ошибка на тесте по несоединенным вершинам незначительно выше, а по соединенным вершинам – незначительно ниже. Из этого следует, что модель не переобучена.

Попытаемся увеличить прогнозную точность при помощи увеличения числа решающих деревьев в модели (используем обучающую выборку, табл. 5, 6).

Таблица 5

Результат классификации Random Forest на обучающей выборке (*ntree* = 1500, *mtry* = 2)

Оценка ошибки OOB			1,07 %
Матрица ошибок			
	0	1	Ошибка классификации
0	2400	6	0,25 %
1	37	1562	2,31 %

Анализируя табл. 6 с OOB-ошибками (*out – of – bag*) и ошибки прогнозов по категориям, а также график зависимости (рис. 4), можем сделать вывод о том, что наименьшее число деревьев, которое демонстрирует минимум сразу всех трех ошибок, равно 400.

Таблица 6

Зависимость ошибки прогноза от числа решающих деревьев[1]

ntree	OOB-ошибка	Ошибка 0	Ошибка 1
100	1,1 %	0,37 %	2,19 %
200	1,07 %	0, 25 %	2,31 %
300	1,07 %	0, 25 %	2,31 %
400	1,05 %	0,25 %	2,25 %
500	1,07 %	0,25 %	2,31 %
600	1,05 %	0,25 %	2,25 %
700	1,07 %	0,25 %	2,31 %
800	1,07 %	0,25 %	2,31 %
900	1,07 %	0,25 %	2,31 %
1000	1,05 %	0,25 %	2,25 %
1100	1,05 %	0,25 %	2,25 %
1200	1,05 %	0,25 %	2,25 %
1300	1,05 %	0,25 %	2,25 %
1400	1,05 %	0,25 %	2,25 %
1500	1,05 %	0,25 %	2,25 %

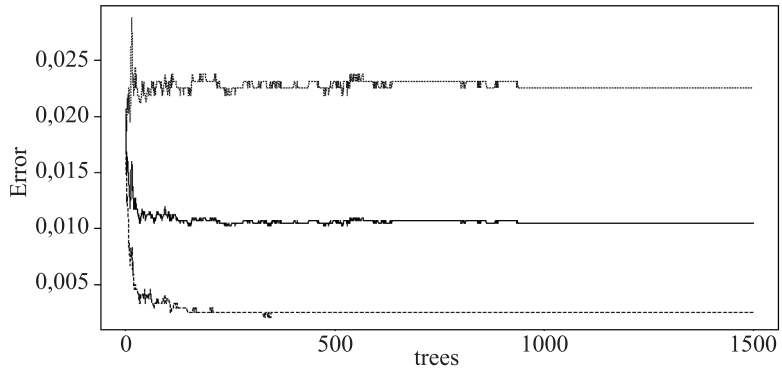

Рис. 4. График зависимости прогнозной точности модели от числа решающих деревьев

Кроме варьирования числа решающих деревьев, качество модели можно улучшить, изменяя количество предикторов, отбираемых на каждом шаге классификатором (табл. 7, рис. 5).

Таблица 7

Прогнозная точность в зависимости от числа отбираемых предикторов

mtry	OOB-ошибка
1	2,49 %
2	1,22 %
4	1,09 %
6	1,39 %

[1] В табл. 6 «Ошибка 0» – ошибка прогноза несоединенных пар вершин, «Ошибка 1» – ошибка прогноза соединенных пар вершин.

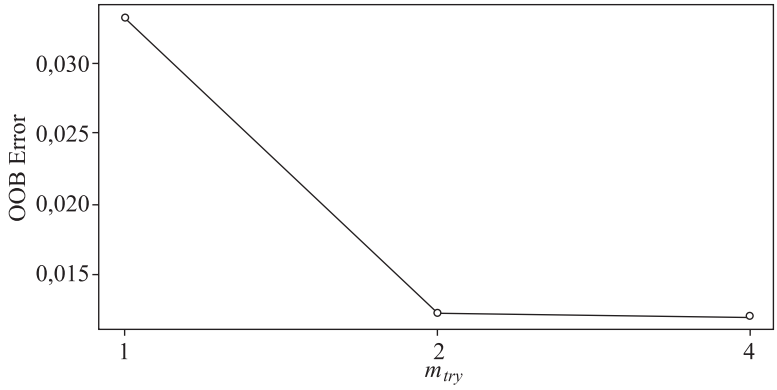

Рис. 5. График зависимости прогнозной точности от числа отбираемых независимых переменных (*mtry*)

Минимум OOB-ошибки достигается при *mtry* = 4. Отметим, что это число отличается от значения *mtry* = 2, выбранного в ходе расчетов автоматически.

На основе проведенного анализа, можно выдвинуть гипотезу о том, что прогнозная точность модели будет выше, если в качестве параметров для классификатора использовать 400 деревьев и 4 предиктора, отбираемых на каждом шаге. Проверим данную гипотезу, построив модель с этими входными параметрами (табл. 8).

Таблица 8

Результат классификации Random Forest на обучающей выборке (*ntree* = 400, *mtry* = 4)

Оценка ошибки OOB			1,07 %
Матрица ошибок			
	0	1	Ошибка классификации
0	2400	6	0,25 %
1	37	1562	2,31 %

Значительно улучшить качество модели не удалось. Построим модель с теми же параметрами на тестовой выборке (табл. 9).

Таблица 9

Результат классификации Random Forest на тестовой выборке (*ntree* = 400, *mtry* = 4)

Оценка ошибки OOB			1,14 %
Матрица ошибок			
	0	1	Ошибка классификации
0	1042	4	0,38 %
1	16	686	2,27 %

Теперь проанализируем важность независимых переменных с позиции влияния на зависимую переменную (табл. 10, рис. 6).

Таблица 10

Значение показателя уменьшения индекса Джини по предикторам

Предиктор	Уменьшение индекса Джини
SP_{t-1}	1661,24
JC_{t-1}	20,46
CC_{t-1}	174,17
$SP_{t-1} - SP_{t-2}$	1,75
$JC_{t-1} - JC_{t-2}$	30,2
$CC_{t-1} - CC_{t-2}$	25,23

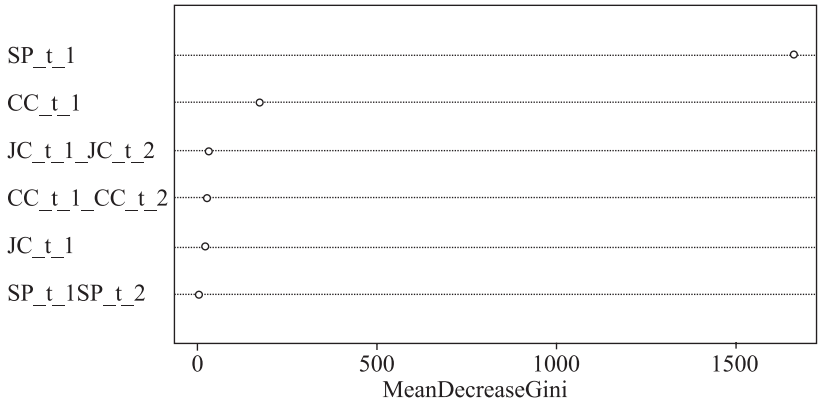

Рис. 6. График динамики уменьшения индекса Джини по независимым переменным

Чем выше показатель уменьшения индекса Джини, тем более значимым является предиктор. В нашем случае самым сильным предиктором оказался показатель кратчайшего расстояния между вершинами, а также суммарный коэффициент кластеризации. Эти результаты полностью совпадают с жизненными реалиями, так как в социальной сети, а особенно на ранней стадии ее развития происходит знакомство либо с «реальными друзьями» (т.е. с теми пользователями, которые и так являются вашими друзьями в реальной жизни), либо с теми пользователями, до которых можно быстро добраться со страниц друзей (т.е. чем меньше длина пути из кликов до пользователя, тем больше вероятность того, что с ним будет установлен контакт).

Малую значимость всех трех разностей $[JC_{t-1}(i, j) - JC_{t-2}(i, j)]$, $[SP_{t-1}(i, j) - SP_{t-2}(i, j)]$ и $[CC_{t-1}(i, j) - CC_{t-2}(i, j)]$ можно объяснить тем, что мы берем для изучения сеть на ее начальном этапе роста, в связи с чем все эти разности являются небольшими. Не исключено, что для уже сформировавшихся сетей эти метрики окажутся значимыми.

Таким образом, мы получим модель, которая способна предсказывать вероятность появления связи между вершинами в последующие моменты времени (также можно предсказывать не только в момент времени $T + 1$, но и в $T + 2$, имея предсказанные значения для $T + 1$). Исследователь, глядя на полученные результаты, может сам решить, при какой вероятности утверждать, что связь между вершинами возникнет, а при какой – нет (стандартная пороговая вероятность равна 0,5).

3.3. Распространение инфекции в сети Flickr

Теперь, имея классификатор, способный предсказывать состояние сети в будущем, попытаемся понять, как его можно использовать для прогнозирования распространения инфекции в сети.

Возьмем экономический пример. Допустим, мы производим смартфоны с камерой 20 МП. Девиз нашего продукта «С новым смартфоном каждый сможет стать профессиональным фотографом». На ранней стадии продаж возникает вопрос: какие каналы распространения мы можем использовать помимо стандартных, проверенных временами?

Если у нас есть партия, допустим, из 30 штук, которые мы можем раздать бесплатно, чтобы запустить «сарафанное радио», то возникает идея воспользоваться сетью Flickr.

Ведь если мы раздадим эти бесплатные телефоны пользователям сети при условии, что они будут делать с помощью наших смартфонов фотографии и выкладывать их в сеть с хэштэгом названия нашего смартфона, то фактически мы запустим в сеть вирус (своего рода вирусную рекламу).

Учитывая то, что у нас есть «здоровые» участники сети (susceptible), а также изначально «инфицированные» нашим смартфоном (infected), то модель распространения нашего вируса по сети может быть описана моделью распространения эпидемии SI (с определенной вероятностью заражения, т.е. мерой того, насколько человек восприимчив к рекламе).

В данной работе не ставится цель рассчитать, каким именно людям мы должны раздать эти самые бесплатные смартфоны, чтобы как можно быстрее распространить инфекцию по сети.

Посмотрим, как может пойти инфекция по сети. Для этого изобразим предсказанную сеть G_{25} с использованием построенной нами модели Random Forest (рис. 7).

Несмотря на изобилие связей, граф не является связным. Выделим в нем гигантскую связную компоненту (рис. 8).

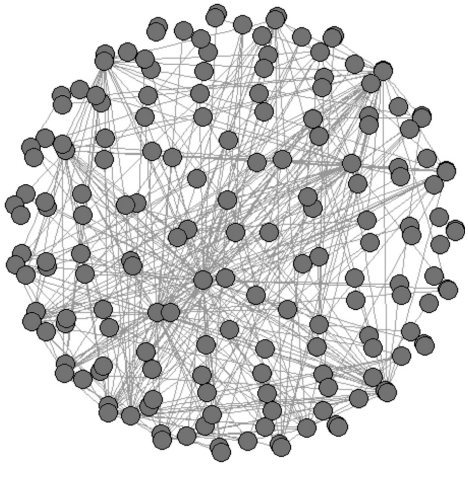

Рис. 7. Граф G_{25}, построенный по предсказанным значениям

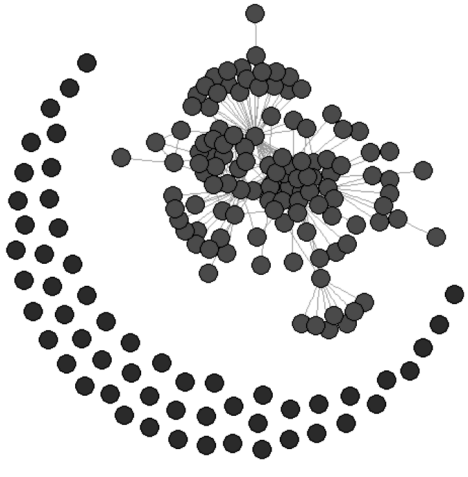

Рис. 8. Граф G_{25} с выделенной гигантской связной компонентой

Красным цветом обозначены вершины, которые входят в гигантскую связную компоненту. Соответственно задачу менеджера можно свести к наблюдению за итеративным изменением сети (т.е. шаг за шагом с выбранным временным интервалом происходит предсказание состояния сети в следующий момент времени, основываясь на текущей информации). Таким образом, заразив какое-то исходное число пользователей в гигантской связной компоненте, можно шаг за шагом наблюдать, как будет идти распространение инфекции. И в силу особенности модели SI, любая вершина, которая на *n*-м шаге присоединится к гигантской связной компоненте, рано или поздно окажется инфицированной.

ЗАКЛЮЧЕНИЕ

В ходе исследования был проведен анализ метрик, используемых для предсказания связей. Как видно из результатов исследования [3], наилучшими метриками являются коэффициенты Катца и Адамика/Адара. Однако в силу того, что расчеты данных показателей на практике оказались довольно затруднительными (нехватка памяти и большое количество времени, затрачиваемое программой при расчете), в проведенном исследовании вышеупомянутые метрики были заменены на длину кратчайшего пути между вершинами и коэффициент Жаккарда соответственно.

Были систематизированы существующие модели предсказания связей в социальных сетях. Для случая классификации по типу используемых при анализе данных были выделены *статический* и *динамический* методы. В силу того что метод статического прогнозирования связей в сети не учитывает изменения топологии сети во времени, для авторского исследования был выбран метод анализа с использованием временного ряда графов. Соответственно дабы подчеркнуть значимость информации о сети, накопленной к текущему моменту времени, автором было предложено использование характерных коэффициентов (а именно разности метрик за предыдущие промежутки времени).

При выборе инструментария для реализации динамического метода были рассмотрены подход, основанный на схожести вершин (*similarity – based approach*), и подход, основанный на обучении (*learning – based approach*). Последний был признан более практичным подходом (и соответственно был выбран автором) в сравнении с *similarity – based approach*, так как метод оценки схожести вершин лишь ранжирует посчитанные коэффициенты и определение того, в каком месте провести линию, выше которой находятся пары вершин, соединение которых наиболее вероятно, остается на усмотрение исследователя.

С целью определения того, какой метод машинного обучения делает в среднем более точные прогнозы связей в сетях, была изучена статья [1], из которой следовало, что метод случайных лесов (Random Forest) является наиболее оптимальным.

Для решения проблемы предсказания связей на практике была выбрана сеть Flickr. Был построен классификатор Random Forest (с параметрами: число деревьев – 400, число предикторов, отбираемых на каждом шаге, 4), который показал OOB-ошибку, равную 1,05 %, ошибку классификации

несоединенных узлов – 0,25 %, ошибку классификации – 2,25 %. Таким образом, классификатор имеет крайне высокую точность. Хотя, стоит признать, что ошибка является заниженной, так как в рассматриваемом временном интервале скорость появления новых связей в сети отстает от скорости появления новых узлов.

Был приведен реальный пример использования предложенного классификатора для экономической задачи распространения информации о новом продукте. Было показано, что при распространении инфекции по сети внешний наблюдатель способен оценить, какие пользователи социальной сети окажутся подвержены заражению, а какие нет. Используя модель распространения эпидемии SI можно утверждать, что любая вершина графа сети, попадающая в гигантскую связную компоненту, рано или поздно окажется заражена.

Комбинация методов предсказания связей в сети и моделей эпидемий с целью прогнозирования распространения инфекции в сети может стать мощным инструментарием. Предложенная методика одновременно учитывает не только характер распространяемой инфекции (изначальное число зараженных, интенсивность заражения и т.д.), но и топологические особенности сети (а именно их изменение во времени) – что крайне важно при итеративном построении модели эпидемии.

Литература

1. *Hasan M.A., Zaki M*. A survey of link prediction in social networks. Social Network Data Analytics, Springer US, 2011. P. 243–275.
2. *Huang Z., Lin D.K.J*. The time-series link prediction problem with applications in communication surveillance. INFORMS Journal on Computing, 2009.
3. *Liben-Nowell D., Kleinberg J.M*. The link prediction problem for social networks. In Proc. of the ACM Intl. Conf. on Inform. and Knowlg. Manage. (CIKM'03), 2003.
4. *Medlock J*. Mathematical modeling of epidemics. University of Washington, 22&24 Vaf 2002.
5. *Nahla M.A., Ling C*. New approaches for link prediction in temporal social networks, Computer Modeling & New Technologies. 2014. P. 87–94.
6. *Potgieter A., April K., Cooke R., Osunmakinde I*. Temporality in Link Prediction – Understanding Social Complexity. ECO. 2009. Vol. 11. No. 1.

Bibliography

1. *Hasan M.A., Zaki M*. A survey of link prediction in social networks. Social Network Data Analytics, Springer US, 2011. P. 243–275.
2. *Huang Z., Lin D.K.J*. The time-series link prediction problem with applications in communication surveillance. INFORMS Journal on Computing, 2009.
3. *Liben-Nowell D., Kleinberg J.M*. The link prediction problem for social networks. In Proc. of the ACM Intl. Conf. on Inform. and Knowlg. Manage. (CIKM'03), 2003.
4. *Medlock J*. Mathematical modeling of epidemics. University of Washington, 22&24 Vaf 2002.
5. *Nahla M.A., Ling C*. New approaches for link prediction in temporal social networks, Computer Modeling & New Technologies. 2014. P. 87–94.
6. *Potgieter A., April K., Cooke R., Osunmakinde I*. Temporality in Link Prediction – Understanding Social Complexity. ECO. 2009. Vol. 11. No. 1.