TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Arvutitehnika instituut

IAF70LT
Artjom Kurapov, 020655

Sõnumite sotsiaalvõrgustikes levimise astmete visualiseerimine jõugraafikute abil

Magistritöö

Juhendaja: Helena Kruus,
Teadur, magister

Tallinn, 2011

ТАЛЛИННСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
Факультет информационных технологий
Институт вычислительной техники

IAF70LT
Артём Курапов, 020655

Визуализация эволюционных каскадов сообщений в социальных сетях с помощью силовых графов

Магистрантская работа

Руководитель: Helena Kruus
Научный сотрудник, магистр

Таллинн, 2011

TALLINN UNIVERSITY OF TECHNOLOGY
Faculty of Information Technology
Department of Computer Engineering

IAF70LT
Artjom Kurapov, 020655

Visualization of social network evolutionary cascades of messages using force-directed graphs

Master thesis

Supervisor: Helena Kruus,
Researcher, MSc

Tallinn, 2011

Autorideklaratsioon

Olen koostanud antud töö iseseisvalt. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt prinevad andmed on viidatud.

Kasutatud Pling.ee andmed on kooskõlastatud Elisa Eesti AS juhatusega ja ei sisalda isikliku identifitseeriva andmeid.

Artjom Kurapov,
17.05.2011

Основные понятия

Граф – совокупность множества вершин (объектов) и рёбер (связей). Представляется графически или матрицей. Основа описания сети.

Каскад – последовательное распространение сообщения (мема, информации, болезни) в сети

Извлечение данных (*Data mining, knowledge discovery*) – процесс обнаружения в сырых (базах) данных, ранее неизвестные, интерпретируемые и практически полезные закономерности для человеческой деятельности [**1**]

Степень вершины — число рёбер у выбранной вершины

Мемплекс – объединение информации социальной сети (мемов) в непротиворечивую систему (идеологию, религию, этнический стереотип)

Сокращения

AJAX — *Asynchronous Javascript and XML*, функция браузера делать http запрос по необходимости с асинхронной обработкой результата

JSON — *JavaScript Object Notation*, синтаксис описания объектов в Javascript, использующийся из-за краткости и простоты в качестве формата передачи данных с помощью AJAX

RDF — *Resource Description Framework*, семейство форматов представления семантических данных

FOAF — *Friend of a Friend*, формат семантического описания социальной сети

SIOC — *Semantically-Interlinked Online Communities*, формат описания социальной сети

UI — *User Interface*, (графический) интерфейс пользователя

GCC — *Giant Connected Component*, основная связанная компонента графа

SMS — *Short Message Service*, услуга передачи текстового сообщения по телефонным сетям между абонентами

API — *Application Programm Interface*, общее обозначение программного интерфейса одной службы для внешних служб. Технические детали и протоколы варьируются (SOAP, REST)

IM — *Instant Messenger*, тип программ передающих мгновенные сообщения через интернет. Детали по централизации, групповым конференциям, передачи файлов, аудио и видео-потоков варьируются в зависимости от протокола и программы.

CUDA — *Compute Unified Device Architecture*, принадлежащая компании Nvidia, технология программно-аппаратного параллельного вычисления программ с использованием графических процессоров и памяти (GPU) вместо обычных (CPU)

OpenGL — *Open Graphics Library* и в частности подмножество WebGL. Кроссплатформенная открытая графическая библиотека и API, управляемая Khronos Group.

Обозначения в формулах

n (*nodes, vertices*) – число вершин
e (*connections, edges*) – число рёбер
l (*length*) – геодезическое расстояние между вершинами
d (*diameter*) – диаметр графа
r (*real distance*) – геометрическое расстояние между вершинами

Аннотация

УДК 311.2, 519.17, 004.421

Данная работа описывает визуализацию сетей и междисциплинарный вопрос распространения каскадов. В частности рассматриваются вопросы социологии о распространении и слиянии информации в обществе. При этом кратко описываются необходимые основы теории графов, существующие модели и научные работы. Описываются алгоритмы расстановки вершин и наиболее функциональные и популярные программные средства реализующие их. Далее предлагается рекурсивный алгоритм расстановки вершин и результаты проделанной работы по его реализации на примере проделанных двух небольших выборок данных социальной сети Twitter. Созданный инструмент визуализации работает в интернет браузере. Для сравнения также предлагается анализ более объёмной выборки из социальной сети Pling.ee (40-70 тысячами пользователей) с помощью инструмента визуализации Gephi. Работа может быть полезна как для ознакомительных целей с предметной областью, так и как практическое пособие при выборе или создании средства визуализации сети или исследованиях в смежных темах

Содержание

Список изображений

Список графиков

Список таблиц

Список алгоритмов

1. Введение

Сеть в абстрактном понимании основывается на взаимосвязи элементов. Зависимость характеристик сети от её природы представляет большой научный интерес, из-за прикладного значения получаемой модели.

В частности с моделированием социальных, телекоммуникационных, биологических сетей становятся доступны методы по оптимизации энергии, профилактике уязвимостей, защите от намеренных атак и предсказании развития, не говоря уже о том, что работа с моделью значительно дешевле и быстрей экспериментов с живыми сетями.

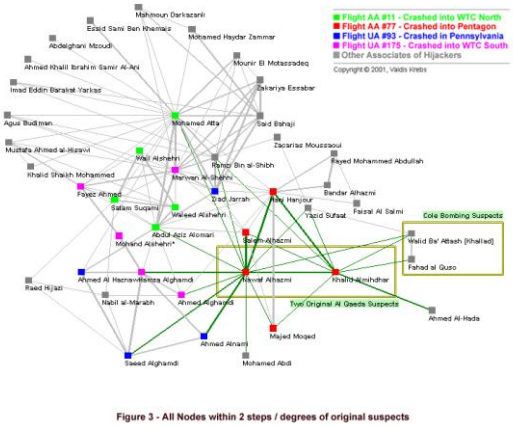


Рисунок 1. Террористическая ячейка, организовывшая взрывы 11 сентября 2001 г. в Нью-Йорке [**91**]

Человеческие социальные связи в частности, с развитием информационных сетей, всё меньше зависят от местных физических структур (транспорта и телекоммуникаций) и всё больше глобализируются и виртуализируются, ускоряя инновации. Вместе с этим растут и риски связанные с политикой, криминалистикой (Рисунок 1), эпидемиологией, телекоммуникацией, финансами.

Человеческий вид единственный использующий торговлю ради улучшения специализации и как следствие уровня жизни. Производство современных товаров массового потребления невозможно без всей цепочки зависимостей между производствами в разных областях техники и географии, более того – никто не в состоянии полностью охватить своим разумом понимание всех тонкостей этого процесса[1]

[1] Например производство монитора зависит от производства пластика, которое зависит от добычи нефти

Общество таким образом можно считать взаимосвязанным коллективным сознанием, генерирующим как новые товары, так и новые идеи об инновациях, событиях, культуре, имеющим свою динамику и закономерности [**2**].

Источниками данных о структуре социальной сети могут служить API популярных социальных сетей (facebook, twitter, linkedin), универсальные децентрализованные семантические сети как *Giant Global Graph* [**3**] на RDF[2] протоколах, данные имеющиеся у государства[3] и разумеется закрытые данные банков и частных компаний

1.1 Объект исследования

Основным феноменом для изучения в таком коллективном сознании является каскад распространения идей, то есть динамического процесса объединения разных тем в единое целое.

Поводом и сутью таких процессов могут быть распространение инфекции [**4**], рекламы, мема[4] или пропаганды. Идея может распространяться и пассивно, без явных сообщений в электронных СМИ как распространение ожирения [**5**], счастья [**6**], коррупции [**7**] или гриппа [**8**]

Своим напряжением наиболее интересна кризисная, бифуркационная ситуация в случае правительственного переворота, падения финансовых рынков, природного катаклизма, террористического акта или эпидемии и тем как общество реагирует на фильтрацию лжи, самоорганизацию и взаимопомощь, или же наоборот, как слепо следует панике (Рисунок 2) или мародёрству.

[2] Имеются ввиду протоколы описания друзей и сообществ FOAF, SIOC

[3] В частности в Эстонии – XTee и Коммерческий регистр данных (E-Äriregister)

[4] Единица культурной информации [**10**5]

Политикам, журналистам и работникам рекламных агентств хорошо известна ситуация, когда громкое событие совпадает с намеченными планами и отвлекает *общественное внимание*, уменьшая резонанс, прибыль, известность.

1.2 Цели работы

В ходе работы исследуется вопросы:

- Как визуализировать сеть с помощью графа?
- Что влияет на появление и распространение каскадов?
- Насколько похожи сети заявленных друзей и действительной сетью общения?

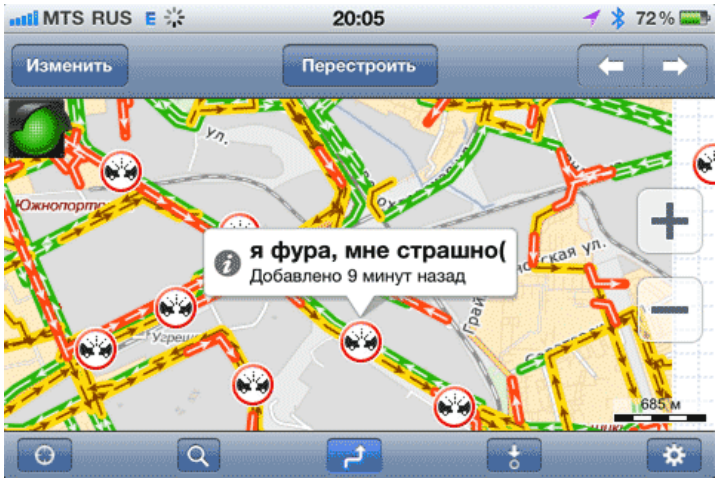


Рисунок 2. Информативность сообщений не всегда та к чему стремились создатели приложения «Яндекс пробки» в iPhone, 11-12 ноября в Москве. [**113**]

2. Теория графов в социологии

В середине XX века психолог Стэнли Милграм провёл несколько экспериментов по изучению эффективного диаметра социальной сети жителей США [**9**]. В результате эксперимент получил общее название «феномен маленького мира» (*small world phenomena*), поскольку полученные данные поражали тем насколько близки люди друг к другу.

В дальнейшем феномен глобализировался в гипотезу об эффективном диаметре социального графа всех людей на планете равному шести (*six degrees of separation).* Данные взятые в 2006 году на основе 242 миллионов аккаунтов из информационной сети «*MSN messenger*» подтвердили эффективный диаметр в 6.6 [**10**]. Дальнейшее изучение сетей с использованием математической теории графов породило целое семейство проблем связанных с рисованием (Рисунок 3), кластеризацией и поиском пути

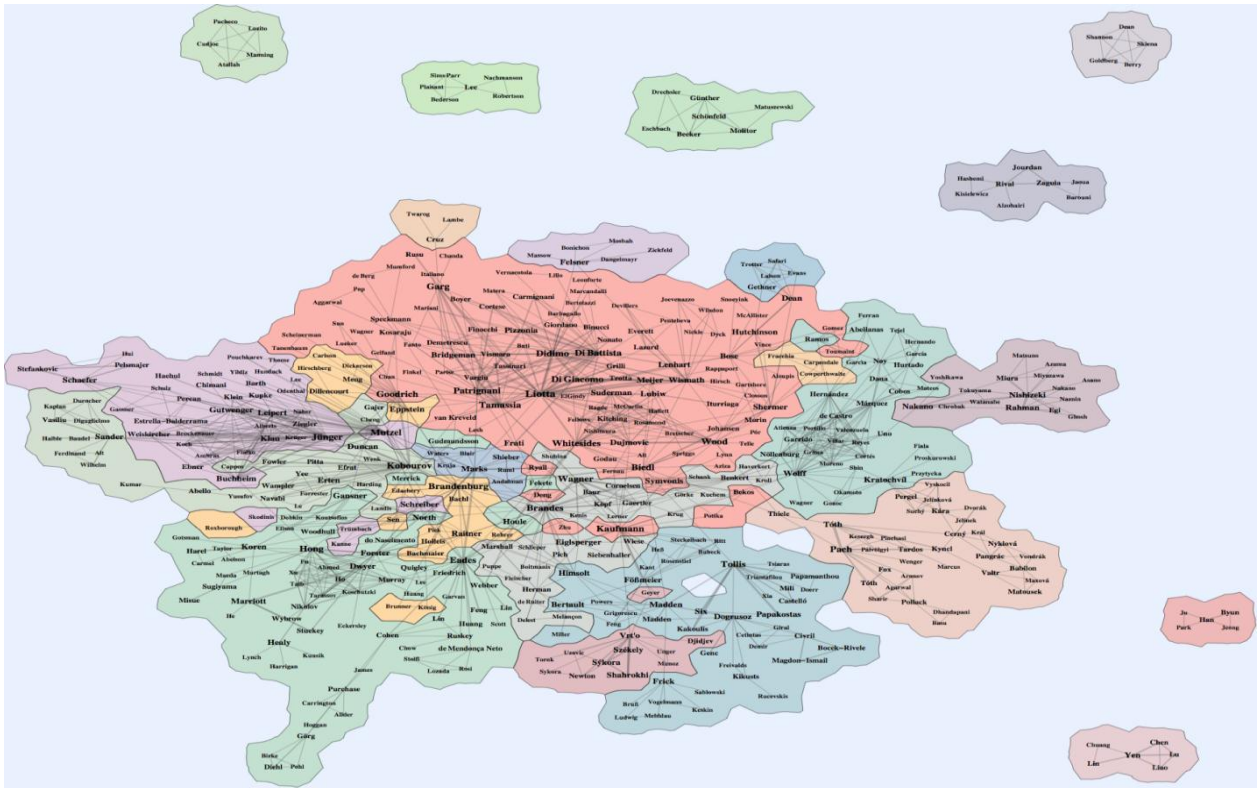


Рисунок 3. Граф учёных, цитировавших друг друга в статьях на тему рисования графов [**11**]

2.1 Типы сетей

Не смотря на различные размеры (Таблица 1), сети выделяются тематически, и наиболее похожие, близки по своим характеристикам. Исследования разных типов сетей и общих характеристик были обобщены [**12**] и выделяются:

- Технологические (электросети, водопровод, почта, транспорт)
- Биологические (нейроны, пищевые цепи, кровеносные сосуды, реки)
- Биохимические (трансформации веществ, влияния ферментов)
- Информационные (цитирования, патенты, лексикон)
- Социальные (друзья, соавторства, сексуальные партнёры, деловые связи)

Сеть	Вершин	Рёбер
Коннектом червя[5]	302	5000
Молекула белка ATN1	5020	5128
Английский словарь [6]	110 000	150 000
Коннектом человека[7]	10^{11}	10^{15}
Звёздные системы во вселенной	$3*10^{22}$	

Таблица 1. Примеры размеров некоторых типов сетей

2.2 Свойства графов в целом

Естественные сети отличаются от случайных графов с фиксированными размерами, предложенных [**13**] и [**14**] а так же деревьев [**15**] поскольку первые постоянно развиваются и как будет показано далее, распределение вероятности связи новой вершины с любой другой вершиной не равномерное.

[5] Имеется ввиду нейронная сеть червя Caenorhabditis elegans, в качестве рёбер приведены химические синапсы [**109**]

[6] Т.н. Wordnet - 70% существительных, 17% прилагательных , 10% глаголов и 3% наречий; 60% Гиперонимов и гипонимов, 30% антонимов, отношений частичности или схожести [**98**]

[7] Только на уровне клеток-нейронов [**110**]

Более того, во многих естественных сетях диаметр графа со временем не увеличивается, а уменьшается, а скорость роста придерживается степенному закону (*power law*) [**16**].

Сеть по масштабу свойств можно условно разделить на три уровня — вершина и ребро, скопление и сеть в целом.

Свойства вершин и рёбер

Общие свойства описываются теорией графов и в зависимости от типа сети могут обладать кроме степени вершины:

- Ориентированностью или замкнутость рёбер
- Дольностью вершин где есть разные их типы — например компании и работники (*multipartite graph*)
- Качественными характеристиками (цвет, масса, форма, температура)

Геодезическое расстояние (*Geodesic distance, shortest path*) — минимальное число промежуточных вершин кратчайшего пути между двумя вершинами [**19**]. Активно используется в маршрутизации TCP/IP пакетов и в играх.

Алгоритмически находится созданием матрицы расстояний с приоритетным индексированием вершин с наименьшими весом [**20**] но в не взвешенных графах сложность алгоритма $O(n^2)$.

Более быстрые алгоритмы имеют сложность $O(n + e * log(e))$ и вводят структуру данных фибоначчиевой кучи [**21**] или кучи остатков (*radix heap*) [**22**] со сложностью $O(e + v * \sqrt{\log(l)})$.

Число Данбара — максимально допустимая степень вершины, в случае с социальными сетями лежит в промежутке 100-230 постоянных связей которые человек в состоянии помнить и поддерживать [**23**].

Близость (*closeness*) — средний кратчайший путь от вершины ко всем другим вершинам графа, показатель относительной центральности.

Промежуточность (*betweenness*) — число присутствия вершины в кратчайших путях между любыми другими вершинами.

Встроенность ребра (*embeddedness*) — число общих вершин у концов ребра. Близкая аппроксимация промежуточности, в социальных сетях используется для оценки вероятности увольнения работника из-за чувства недостатка важности его труда для коллектива

Центральность по собственному значению (*eigenvector centrality*) — рекурсивная характеристика важности вершины получаемая из суммы важности связанных вершин, используется в частности в алгоритме Page Rank

Центральность Маркова — среднее число промежуточных вершин полученных при случайном блуждании от одной вершины к другой [**24**]

2.2.1 Свойства скоплений

Клика (*clique*) – полный подграф, подмножество вершин, каждое из которых связано со всеми остальными вершинами этого подмножества и которое не принадлежит другому клику. Поиск клика – задача NP класса сложности.

Степень кластеризации (транзитивности) — характеристика повышенной вероятности связи между вершинами A↔C, если A↔B и B↔C (друг моего друга — мой друг). Есть несколько вариантов вычисления, из них комбинаторный алгоритм [**25**] для ненаправленных графов определяется формулой:

$$T = \frac{1}{n}\sum_{i}\frac{\mathrm{c}_i}{e_i(e_i-1)}$$

e_i– степень вершины i, если она менее 2, то дробь в сумме не учитывается
c_i – число уникальных пар

Согласованное смешение (*assortative mixing*) – корреляция между заданными типами вершин и вероятностью их связи. Выявляет расовые, языковые, финансовые, половые и прочие предпочтения в социальных связях.

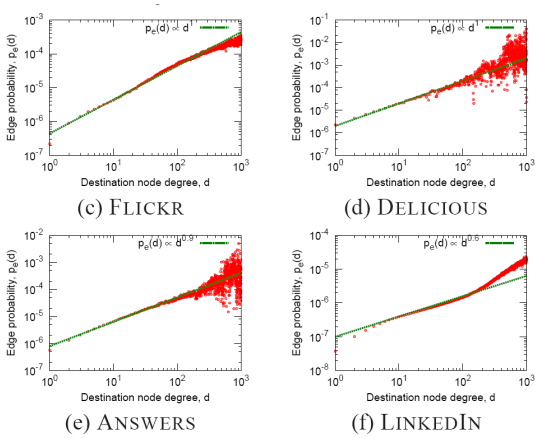
$$Q_{a\leftrightarrow a} = \frac{(\sum_i \frac{e_{i(a)}}{e_i}) - 1}{e - 1}$$

e – число всех рёбер в графе
e_i - степень i-вершины с типом a
$e_{i(a)}$– число рёбер i-вершины соединённых с вершинами класса a

При Q=1 вершины связываются только со своим типом. При Q=0 связь от типа не зависит.

Один из способов описания случайных эволюционирующих графов – модель **предпочтительной связи** (*prefferential attachment*), заключающаяся в предположении, что вероятность создания новой связи зависит от степени целевой вершины [**17**]

Второй вариант стохастической модели [**18**] симулирующей социальную сеть учитывает рождение вершин и создание связей следуя отличной от нуля степени транзитивности



(c) FLICKR (d) DELICIOUS
(e) ANSWERS (f) LINKEDIN

График 1. Предпочтительная связь на примере четырёх социальных сетей [**103**]

2.2.2 Свойства сети

Плотность графа (*density*) – зависимость числа рёбер от числа вершин. В большинстве эволюционирующих сетей зависимость степенная [**18**]

$$e(t) \propto n(t)^a$$

a – степень роста числа рёбер, находится в промежутке между 1 (дерево) и 2 (клика).

Диаметр — максимальный кратчайший путь между любыми двумя вершинами (между которыми такой путь возможно проложить). Нахождение очень затратное по вычислительным мощностям.

$$d = min \max l_{ij}$$

Наибольший связанный компонент (*giant connected component, GCC*) и его диаметр важен в сетях где существуют изолированные друг от друга вершины (между которыми нет транзитивного пути). Диаметр GCC в большинстве естественных сетей максимален при начальной стадии эволюции и со временем уменьшается [**18**]

Средний диаметр (*mean geodesic*) есть среднее гармоническое между кратчайшими путями l_{ij} между вершинами i и j [**12**]. Приблизительное значение можно вычислить выбрав n случайных вершин (а не проходить целиком весь граф) [**26**]

$$d_{ef} = \frac{(n-1) * \frac{n}{2}}{\sum_{i \geq j} \frac{1}{l_{ij}}}$$

Распределение степеней (*degree distribution*) — график зависимости степени вершины от всего количества таких вершин в графе. Из-за своей дискретной природы получается шумная функция и для её сглаживания при больших i, используются увеличивающиеся интервалы

$$p_k = \frac{1}{2^t} \sum_{j=2^t}^{2^{t+1}-1} p_j$$

Минимальный разрез (*minimum cut*) это разделение графа на подграфы с минимальными потерями связей (или их весов). Как правило, именно так разделяется группа в случае социального конфликта между центрами скоплений [**27**]

2.3 Социальные сети

2.3.1 Типы социальных связей

При рисовании графов как социальных сетей, лица (физические и юридические если это двудольный граф) считаются вершинами. Однако интерпретация наличия осмысленных связей и соответственно их кодирование связей в графическом дизайне порой важней формальных вершин-участников. Именно смысл, осознание и намерение оценивается в юридической практике при оценке умышленности деяния.

Движение в толпе, просмотр парада, действие хирурга над пациентом и даже пулемётный расстрел как нацистами [**28**] так и солдатами США [**29**] — не социальные связи, поскольку для социального поведения необходимо сопереживание, т.е. связь вершин одного рода, а не механическая работа над неодушевлённым объектом.

Социологи [**30**] выделяют следующие типы социальных связей по мере усложнения:

- Поведение — направлено в отношении конкретного человека (разговор, агрессия, альтруизм, застенчивость, изгнание и жертвенность)
- Действие – происходит с учётом поведения окружающих. Делятся на:
    - Целе-рациональные, на основе имеющихся ресурсов (управление компанией, конкуренция, борьба, обмен)
    - Ценностно-рациональные, на основе веры и ценности, независимо от побочных последствий (борьба Ганди; дружба, благотворительность)
    - Традиционные – на основе устоявшихся шаблонов поведения (приветствие, совместный обед, ложь)
    - Эмоциональные (любовь, игра)
- Контакт — знакомство, создание новых отношений
- Взаимодействие (влияние) [8] — поведение или действие группы которое вынуждает индивида к определённому участию (переписка с другом, беседа с громкими

[8] Social Interactions

соседями, планирование военной доктрины, игнорирование работы конкурента) независимо от физического расстояния или типа отношений но непременно с взаимной ориентацией.

- Повторяющееся взаимодействие (случайное, незапланированное)
- Регулярное взаимодействие (запланированное)
- Регулируемое взаимодействие (на уровне всей сети – закона, традиций, стандарта)

2.3.2 Распространение каскадов в социальной сети

Каскад — процесс распространения **данных** в графе в зависимости от времени, который можно представить в виде направленного дерева. В контексте исследуемой сети twitter это массовый retweet сообщения который в можно обобщить до **общественной мысли**, который закрепляется практическими примерами из жизни – мемах.

На индивидуальном уровне, каскад может представляться советом при покупке, копированием **социального поведения** (битьё окон при протесте, мода на юбки, курение) или в частности передаче информации по цепочке.

При распространении каскада сообщений, изначальные данные видоизменяются как из-за энтропии, так в результате естественной эволюции, т.н. **биссоциации** [**31**] — слияний и мутаций с другими сообщениями, теряя изначальный смысл и угасая по мере распространения.

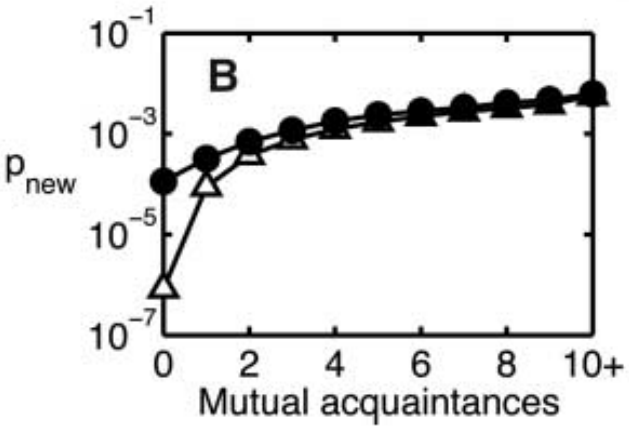


График 2. Кривая зависимости между числом общих друзей и вероятности новой связи в анализе рассылки почты [**104**]

Идеологию в таком случае можно определить как скопление взаимодополняющих идей у конкретной группы.

Кривая диффузии — это показатель скорости распространения сообщения, вероятность принятия социального поведения у вершины в зависимости от относительного числа общих вершин с этим поведением и силами связей с ними. [**32**]

В некоторых каскадных процессах можно наблюдать кривую подобно убывающей доходности, как у передаточной функции искусственного нейрона.

Например, формула распространения инноваций [**33**]:

$$N_t = N_{t-1} + p(m - N_t - 1) + q\frac{N_{t-1}}{m}(m - N_{t-1})$$

m – потенциал рынка (максимальное число охватываемых вершин),
p – коэффициент внешнего влияния (масс-медиа; можно считать общим распространением в сети), в предлагаемом примере = 0.03
q – коэффициент подражания (влияние от друзей), в предлагаемом примере = 0.38

Мотив (*network motif*) — направленный подграф с определённой структурой. Можно считать его закономерным распространением каскада (за фиксированное время). Активно изучается в экологических и биохимических сетях.

Точка перехода – состояние системы, после которого начинается массовая реакция, в том числе изменение структуры сети (заражения, переезда, беспорядков, фазы вещества)

Например, для модели заражения с двумя состояниями вершины (болен или здоров) существует формула [**34**] :

$$\frac{\beta}{\delta} < \tau = 1/\lambda_{1,A}$$

τ — эпидемическая точка перехода
$\lambda_{1,A}$ — максимальное собственное значение матрицы смежности
β — вероятности заражения
δ – вероятности излечения заражённого

Диффузия, по-видимому, напрямую зависит от сильных связей, т.е. если η друзей заражены и связаны между собой, то вероятность заразить общую вершину больше, чем

вероятность получения заражения от η несвязанных между собой друзей - социальный капитал [**35**] генерирует больше каскадов.

2.3.3 Слабые силы

Если определить силу взаимосвязи между людьми согласно частоте социального поведения за последний год так что хотя бы дважды в неделю – сильное, хотя бы раз в год – слабое и в промежутке - среднее) то при поиске работодателя, нахождение результата максимально эффективней через средние (55.7%) и слабые (27.6%) связи [**36**]. Это объясняется большей промежуточностью слабых связей между разными скоплениями.

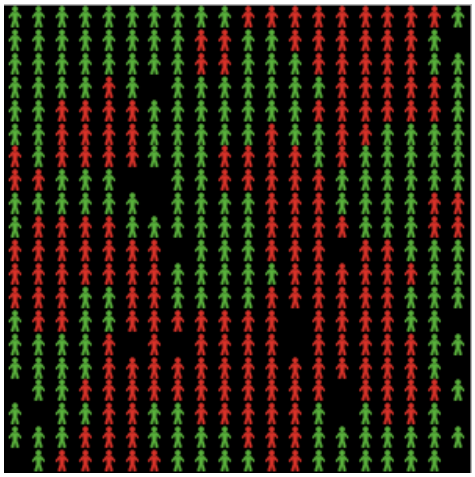


Рисунок 4. Результат пространственной сегрегации Шеллинга при минимуме 31% одинаковых соседей [**111**]

Очень интересна модель пространственной сегрегации [**37**] в которой рассматривается двумерное пространство с жителями двух классов с самоорганизацией.

Житель считается недовольным, если вокруг него находится некое пороговое число жителей другого класса (т.е. сумма отталкивающих сил достигает критической отметки).

Недовольные переезжают на случайную свободную клетку в пространстве. Уже при желании видеть хотя-бы треть жителей такого же класса, как и он сам, возникают массивные кластеры .

Интерпретировать результаты в зависимости от контекста можно как отрицательно (расовое, возрастное, языковое, религиозное и проч. разделение общества), так и положительно (минимизация энергии, максимизация *встроенных связей*)

2.3.4 Важность среды

При анализе распространения информации в социальной сети, недостаточно рассматривать только топологическую структуру, но следует учитывать и обстоятельства, в которых находится источник.

В физическом пространстве человек ограничен возможностью слышать определённое число людей одновременно, вынужден больше обращать внимания на тех, кто кричит громче или находится ближе.

В виртуальных инфосетях меньше таких ограничения, но и тут действуют силы, которые не столько тормозят, сколько искажают каскады. В частности Facebook, Google и другие популярные сайты предоставляют поток результатов не только в персонализированной, но и в отфильтрованной форме, согласно предыдущим предпочтениям читателя [**38**].

Это значит, что сообщения в ленте становятся тематически более близкими и не отражают всего спектра поступаемых сообщений. Тематически близкие сообщения помогают лучше разобраться в теме и углубиться в кластер, но из-за обрезания слабых связей усугубляется изолированность кластеров.

2.3.5 Общественная память

Как было показано [**39**], группы людей формируют память, при этом интеллект группы не коррелирует с интеллектом индивидуумов, а зависит и от числа проводников — более чувствительных коллег, женщин.

Описанная выше фильтрация, таким образом, делает группу более узкоспециализированной и в общем, глупее. Для более гибкой группы, необходимы разносторонние личности, независимость мышления и возможность агрегировать свои идеи [**40**]

2.4 Рисование графа

2.4.1 Критерии

Под построением графа мы подразумеваем алгоритмическую обработку данных и такое визуальное их представление, которое *наглядно* показывает агрегированную структуру сети человеку. Построение графа так или иначе зависит от контекста решаемой задачи - недостаточно случайным образом расположить вершины и соединить их.

В разных приложениях имеет смысл намеренно связывать дизайн с физическим местоположением (для карт), течением времени (для генеалогий), минимизировать длину рёбер и их пересечение, допускать изгибы или ортогональное направление, иметь форму у вершин (в интегральных схемах) [**41**]

В данной работе мы будем использовать критерии для построения **информационных и социальных сетей**. Таким образом основные критерии которые граф должен максимально хорошо учитывать:

1. Уменьшение длины рёбер
2. Увеличение углов между рёбрами
3. Уменьшение пересечений рёбер (это например критично в дизайне интегральных схем для уменьшения числа слоёв)
4. Уменьшение площади

2.4.2 Алгоритмы

Приведённые выше критерии могут друг другу противоречить. Например уменьшение пересечений рёбер и увеличение углов как правило увеличивает площадь и длину рёбер. Для решения тех или иных проблем связанных со скоростью, наборами критериев были предложены многие алгоритмы

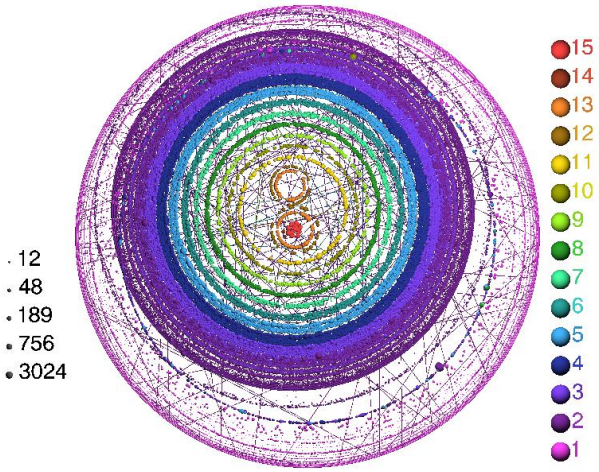


Рисунок 5. Миллион сайтов домена .fr с ядерным расположением и двумя независимыми компонентами на 12 уровне (посередине) [**44**]

- **Уплощение непланарных** графов заключается в удалении рёбер, при котором граф становится планарным. Удаляемые рёбра выбираются так что-бы либо минимизировать их число - (NP-полное время), либо минимизировать число наименее пересекающихся (NP-сложное время). Предполагается что некоторая версия графа к этому времени уже имеется и мы его лишь упрощаем визуально
- Метод **расслоения** графа минимизирует общую площадь и число пересечений [**42**]
- Методы течений максимизируют углы между рёбрами и минимизируют изгибы рёбер
- Упрощение до **физической системы** с силами притяжения и отталкивания с минимизацией общей энергии системы для получения оптимального отображения. Относительно просто реализуется и хорошо смотрится [**43**]
- Ядерная декомпозиция по степени вершины [**44**] (Рисунок 5)

2.4.3 Силовые итеративные алгоритмы

Модель силовых графов во многом опирается и коррелирует с фундаментальными основами и вопросами космологии — отсюда и связь с физикой [9] и трёхмерным

[9] Имеется ввиду теория многих тел в квантовой механике и гравитационная задача N тел в астродинамике

моделированием. Общая идея в том, что бы достичь энергетического равновесия благодаря балансу сил отталкивания и притяжения, подобно физическим законам, которые в свою очередь постулируют зависимость силы от расстояния и характеристик вершин и рёбер

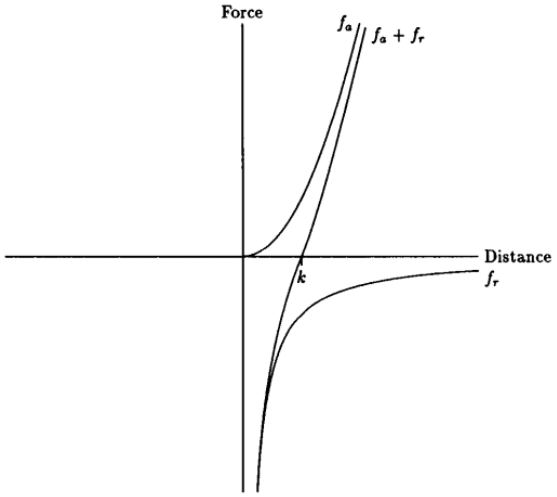


График 3. Сила взаимодействия двух вершин в зависимости от расстояния [**46**]

Силы могут вычисляться как

- Линейной зависимости от расстояния — закон упругости Гука: $F = kr$
- Обратный квадрат от расстояний — законы всемирного тяготения (Ньютона) и взаимодействия электрических зарядов (Кулона):

$$\mathrm{F} = \mathrm{k}\frac{\mathrm{q}_1 \times \mathrm{q}_2}{\mathrm{r}^2}$$

- Логарифмическая зависимость: $F = k\log(r)$

Обычно используется комбинация отталкивающей электростатической силы и притягивающей силой упругости [**45**], впрочем существуют успешные алгоритмы с более сложным вычислением сил [**46**] (График 3. Сила взаимодействия двух вершин в зависимости от расстоянияГрафик 3):

$$f_a(r) = \frac{r^2}{k};\quad f_r(r) = \frac{-k^2}{r}$$

r- расстояние между вершинами,

$k = C\sqrt{\frac{\text{ширина} \times \text{высота холста}}{\text{степень вершины}}}$,

C – подобранный коэффициент

Кроме явного задания сил основанных на геометрическом расстоянии между двумя вершинами ссуществуют алгоритмы:

- Барицентрический [**47**] заключается в использовании линейной зависимости между обоими силами и расстоянием, а что-бы начался процесс балансировки,

некоторые из вершин образуют внешний многоугольник. Минусы — работает только с небольшими (менее 100 вершин) планарными графами с максимальной степенью вершины равной 3

- Пропорциональности сил между вершинами их геодезическому расстоянию [**48**]
- Дополнительных «магнитных» силах влияющих на все рёбра [**49**] и создающих торсионные силы

2.4.4 Поиск глобальных минимумов

В природе так же можно заметить более высокие признаки организации данных которые граф мог бы более явно подчёркивать, такие как симметрия, кластеризация, многомерность, самоподобие. Поскольку алгоритм силовых графов основан на локальной минимизации энергии для конкретной вершины, то он не может минимизировать пересечение рёбер в целом

Для совмещения разных эстетических требований можно использовать алгоритм имитации отжига (*simulated annealing*) [**50**] включающий оценочную линейную функцию эстетической красоты

$$\eta = \lambda_1\eta_1 + \lambda_2\eta_2 + \lambda_3\eta_3 + \lambda_4\eta_4$$

η_i— силы притяжения (квадратичная), отталкивания(обратно квадратичная от расстояния) в том числе и от стенок холста и число пересечений
λ_i- подобранные коэффициенты

Алгоритм заключается в многоразовом нагревании и остывании вершин (подобно закалению стали). В результате находится наиболее прочное состояние с глобальным минимумом энергии

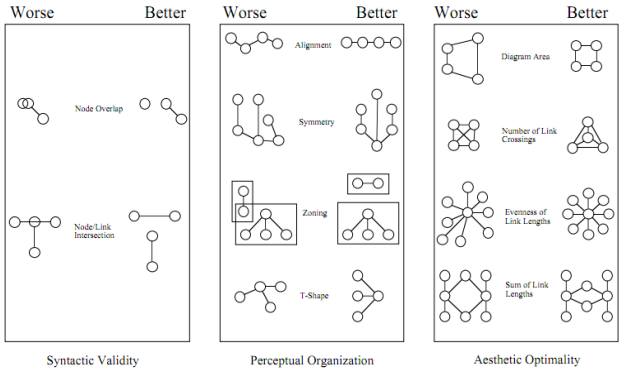


Рисунок 6. Выборы в расположении вершин при обучении в генетическом алгоритме с заданием синтаксической правильности, предпочтении в восприятии и эстетической оптимальности соответсвенно [**85**]

Использование общей многопараметричной

функции энергии так же производилось и с генетическими алгоритмами [**51**] (Рисунок 6)

2.4.5 Уровневое масштабирование

Основная проблема силовых графов состоит в нелинейной сложности, из-за того что движение масс на высоком «галактическом» уровне напрямую рассчитывается для всех вершин.

Многомасштабные алгоритмы (*multiscale methods*) некоторым образом группируют множества вершин и уже между группами вычисляют силы для экономии вычислительных мощностей. Группировка может происходить как в евклидовом пространстве, так и по кластерам, фактически симулируя физику твёрдого тела с центрами масс

Скопления можно находить как на одном уровне (*partitional clustering*) так и вкладывать уровни друг в друга (*hierarchical clustering*), вопрос лишь в том, где и как чётко стоит проводить черту между независимыми скоплениями.

Как показывает практика на примере реализаций в **Gephi**, самый быстрый алгоритм учитывает описанный принцип [**52**]

2.5 Классификация и кластеризация

Классификация необходима для соотнесения элементов множества к выбранным человеком группам. Например определение съедобности фрукта по его цвету и форме. Некоторые элементарные алгоритмы:

1. Перцептрон (пытается провести линию между двумя группами данных)
2. Метод опорных векторов (*support vector machine*) (пытается определить мерность пространства и гиперплоскость для разделения данных)

Кластеризация же наоборот, пытается выявить группы из однородных элементов

2.5.1 Алгоритмы иерархической кластеризации

В данной работе нас будет интересовать иерархическая и интеграционная кластеризация (*hierarchical agglomerative clustering*), которая заключающаяся в объединении вершин в кластеры с нижних уровней (в отличие от разделительных алгоритмов начинающих с высоких уровней и продвигающихся вниз)

Некоторые типы:

- По степени транзитивности (*nearest neighbour*)
- По центру масс (Дисперсионные, *variance*, *k-means*) – число скоплений надо знать заранее, неизвестно как заранее выбирать исходные центры масс

3. Близкие работы

Область визуализации и анализа сетей очень обширная и нельзя не включить опыт и выдающиеся практические результаты коллег, как из научной, так и из коммерческой области.

3.1 Программные продукты

На рынке существует масса программ способных как рисовать, так и анализировать данные любых сетей. Приведём краткий обзор возможностей некоторых некоммерческих продуктов.

3.1.1 Pajek

Программа написана на Delphi, отличается способностью визуализации 10-100 тыс. вершин с различными способами расположения и с алгоритмами поиска закономерностей. Кроме того поддерживаются двудольные графы, параллельные рёбра, развитие во времени.

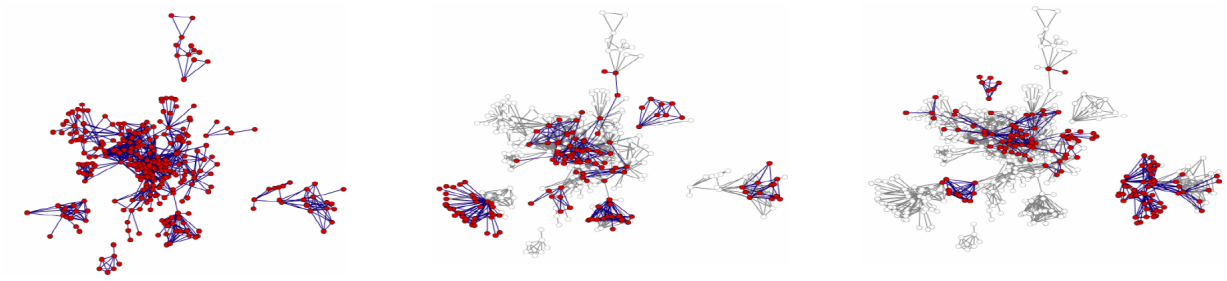


Рисунок 7. Эволюция сети наркоманов в Colorado Springs [**53**] показана с помощью Pajek (в данном случае в течение трёх лет)

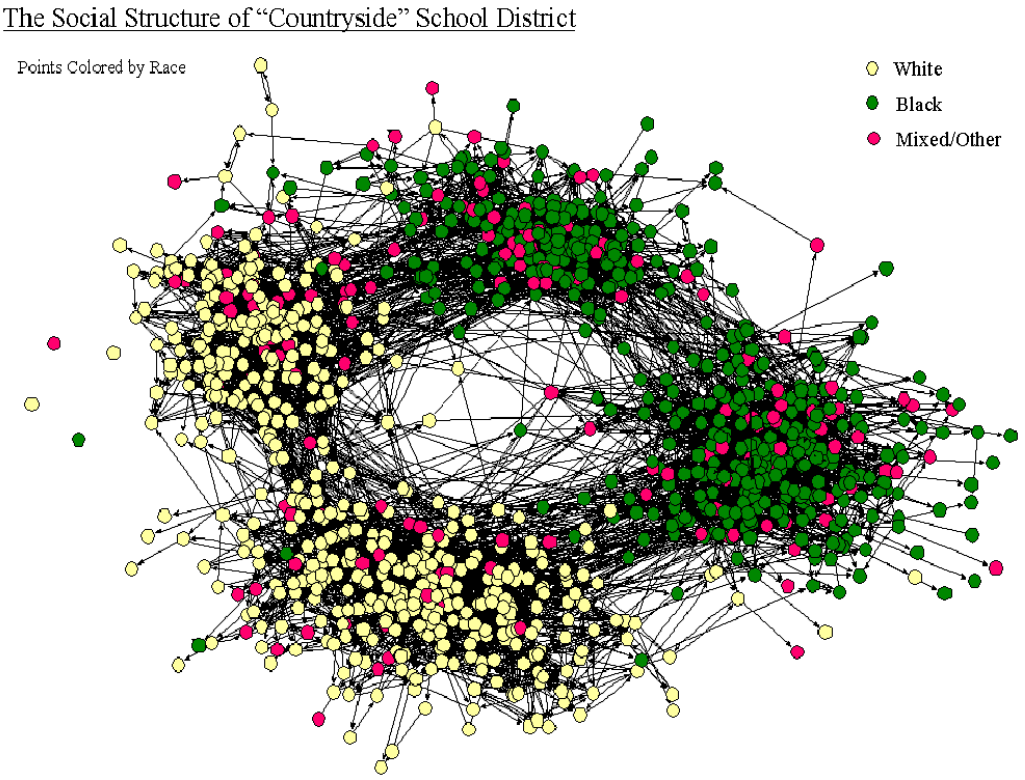


Рисунок 8. Пример согласованного смешения в сети дружбы в школе штата Огайо в зависимости от цвета кожи, зелёным цветом обозначены чернокожие, жёлтым – белые ученики (Moody, 2001) (V., 2005) с помощью Paiek

3.1.2 NetworkX

Написанный на Python поддерживает мультиграфы, направленные рёбра, определение коэффициента кластеризации а также рисование с помощью GraphViz [**54**]

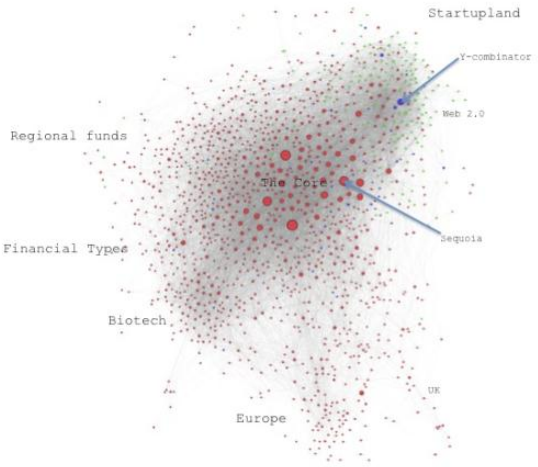


Рисунок 9. Карта венчурных инвесторов на основе данных Crunchbase в NetworkX. [**95**]

3.1.3 SocNetV

SocNetV [**55**] написан на C++ и Qt в основном под Linux платформы с поддержкой различных форматов

импорта данных[10], генератором случайных графов и встроенным web-crawler. Отрисовка поддерживает радиальный, уровневый и энергетические алгоритмы расположения вершин. Анализ поддерживает не только определение основных свойств сети но и расчёт матрицы расстояний.

В частности интересна поддержка различных свойств центральности (промежуточности) - *stress centrality*, *graph centrality*, *eccentricity centrality*, *power centrality* и в частности информационная центральность [**56**], которая важна в данной работе.

3.1.4 NodeXL

Расширение для Microsoft Excel позволяющее импортировать данные из Twitter, Youtube, Flickr кроме обычных GraphML и собственно Excel-данных. Позволяет настраивать раскраску и искать подгруппы с помощью библиотеки SNAP [**57**] на C++

Рисунок 10. Визуализация части сети Twitter по ключевому слову с помощью NodeXL [**93**], в качестве вершин используются аватары

[10] GraphML, GraphViz, Pajek

3.1.5 Graph Tool

Написанный на *python*, этот инструмент тоже кроме стандартных методов ввода-вывода, фильтрации и генерации данных, позволяет запускать методы по анализу скоплений и пропускной способности сети (*flow control*):

- Алгоритм Диница/Эдмондса-Карпа по поиску максимального потока между выбранными вершинами [**58**] [**59**]
- Алгоритм Бойкова-Колмогорова [**60**]
- *Push-relabel* алгоритм [**61**]

Кроме этого, инструмент позволяет работать с мотивами (стр. 11) — есть методы поиска их числа [**62**] а также их относительной важности.

3.1.6 CFinder

Написанная на java программа по поиску сообществ в графах, использует алгоритм перклоляции клика[11] [**63**] который по сути ищет клики внутри всей сети, при этом они могут пересекаться. Программа в основном ориентирована их создание списка выявленных сообществ и кликов и на их рисование.

3.1.7 Gephi

Написанная на Java и OpenGL на основе модульной NetBeans платформы, эта программа одна из наиболее популярных — позволяет импортировать данные напрямую из базы данных MySQL, поддерживает расширения для разных алгоритмов отрисовки и кластеризации, выдаёт статистику близости, развитие во времени и в 3х-мерном пространстве, позволяет обновлять данные через API в режиме потока.

Gephi поддерживает фильтрацию, кривые Безье, многоуровневый алгоритм позиционирования (*Graph coarsening*). Модули в свою очередь позволяют интеграцию с графовой БД Neo4j, семантической Freebase.com, web-браузером.

[11] Clique percolation method

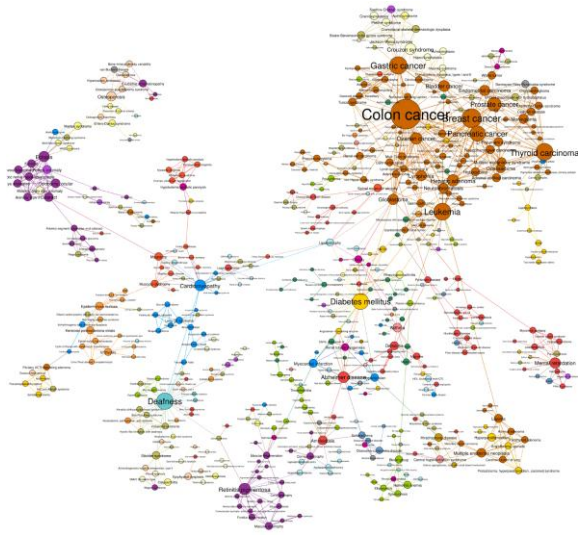


Рисунок 11. Сеть человеческих болезней в Gephi на основе Diseasome, где рак связан с наибольшим скоплением [**101**]

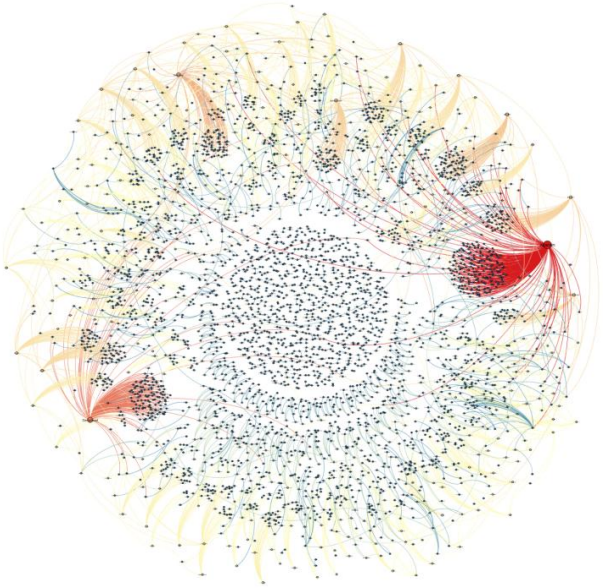


Рисунок 12. Повторение сообщения (Retweet) в сети Twitter с тэгом #jan25, день вынужденной отставки Хосни Мубарака с помощью Gephi [**99**]

3.1.8 Tulip

Написанная на C++ платформа поддерживающая до 1 млн вершин с плагинами дающими просмотр статистики и гистограмм, разными алгоритмами расположения вершин, фильтрацию и создание подграфов, искажение проекции с рыбьим глазом (Fish-eye lens). Наиболее интересна функция представления графа в виде самоорганизующихся карт (self organizing map)

3.2 Коммерческие анализы каскадов

3.2.1 SocialFlow

Компания SocialFlow провела анализ [**64**] распространения новости о смерти Усамы бин Ладена в сети Twitter за два часа до сообщения этой новости президентом.

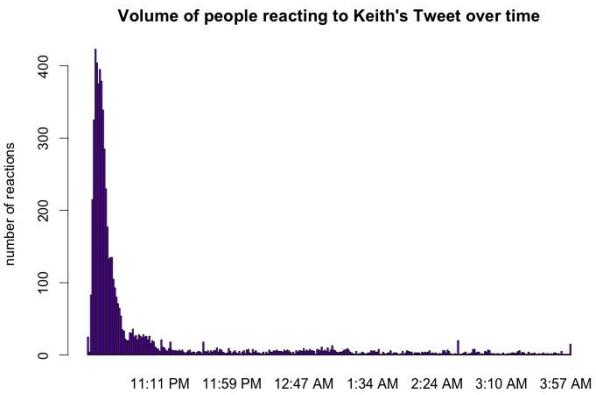


График 4. Распространение Twitter-сообщения пользователя keithurbahn [**64**]

Многие пользователи заранее стали гадать о причине созыва пресс-конференции, оценивать

вероятности, впрочем этого было недостаточно для вирусного эффекта. За час до пресс-конференции сообщение Кита Урбана (*Keith Urbahn*) со ссылкой на неизвестный источник, сообщение вызвало массовый retweet с каскадным распространением через *hub*-вершины

3.2.2 NYTLabs

Отдел разработки газеты New York Times и Марк Хансен[12] сделали инструмент для визуализации распространения сообщений в Twitter [**65**].

Основанный на MongoDB[13] и Processing[14], инструмент рисует трёхмерное ступенчатое радиальное дерево, где горизонтально отмечаются сообщения в зависимости от времени, соединённые с источником, а вертикально – глубина распространения сообщения. Независимые каскады первого уровня в свою очередь группируются по кругу.

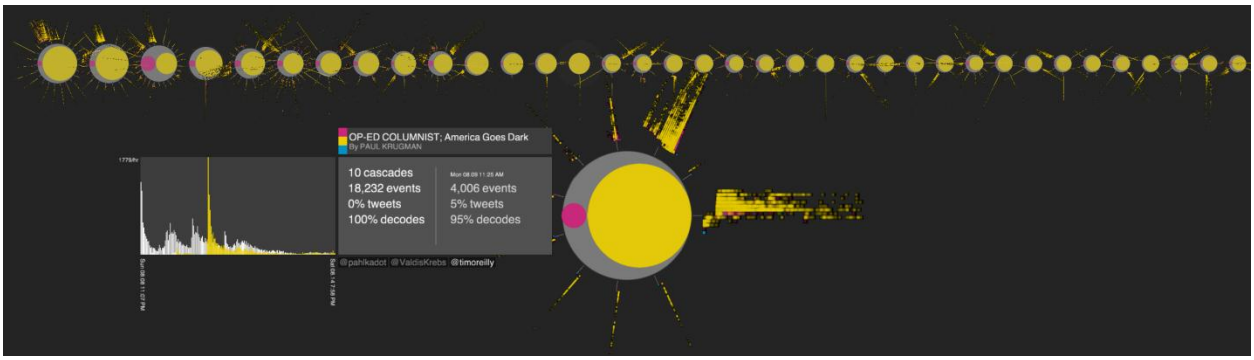


Рисунок 13. Древовидный каскад об отлючении электричества в Америке [**65**]

[12] Mark Hansen, профессор Статистики в UCLA

[13] Систему управления базами данных, не требующая жёсткой типизации (NOSQL), не поддерживающая SQL запрос, масштабируемая репликацией и нуждающаяся в соответсвующем хранении данных для распределённого поиска с алгоритмом map-reduce.

[14] Язык программирования, ориентированный на визуализацию и простоту синтаксиса для незнакомых с программированием людей

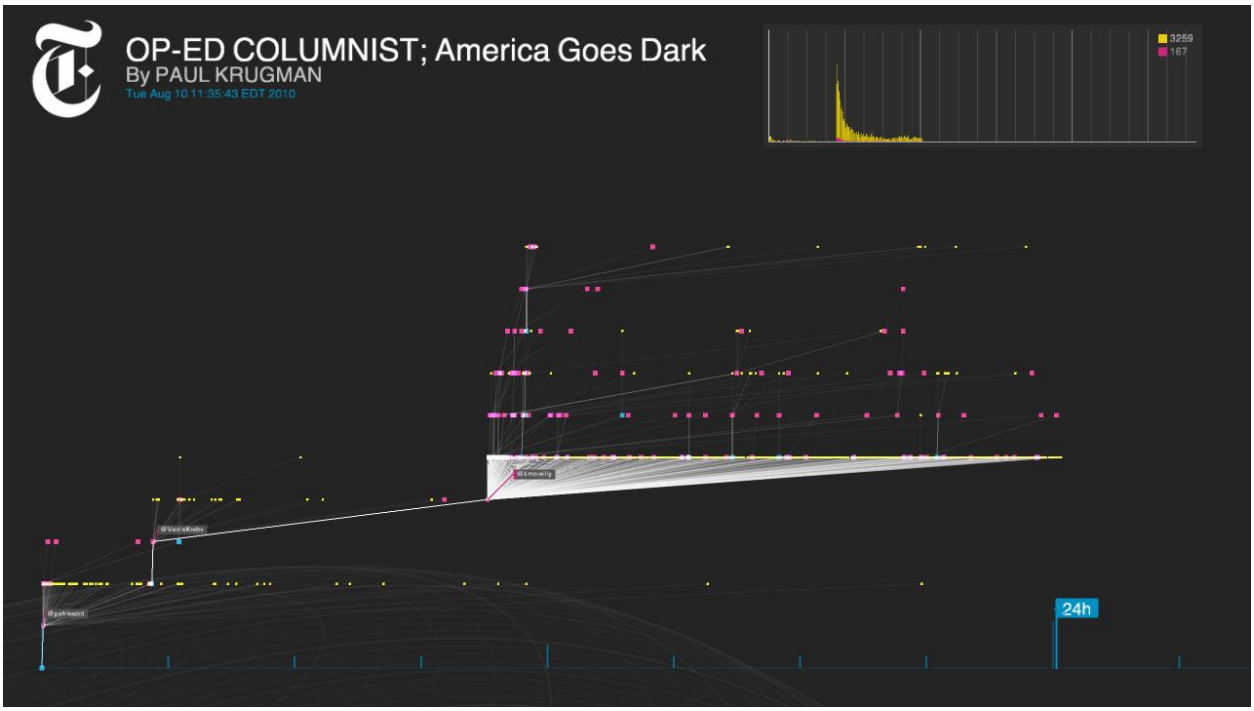


Рисунок 14. Диаграмма относительного распространения сообщений от конкретной вершины [**65**]

3.3 Сайты

3.3.1 Truthy

Truthy— сайт сделанный группой разработчиков Университета Индианы [**66**], на основе Django, MySQL, Gephi Toolkit и Boost, который постоянно следит за развитием мемов в как на основе хэш-тэга [15], так и по ссылкам на внешние ресурсы и перекрёстные упоминания [**67**].

Свою целью, ресурс ставит разоблачения скрытых акций, выявление истинных взаимосвязей между участниками разговоров (Рисунок 16). Добавление объектов мониторинга доступно только авторам.

[15] Хэш-тэг — символьное обозначение некоей семантики в сообщении (темы, события) символом «#»

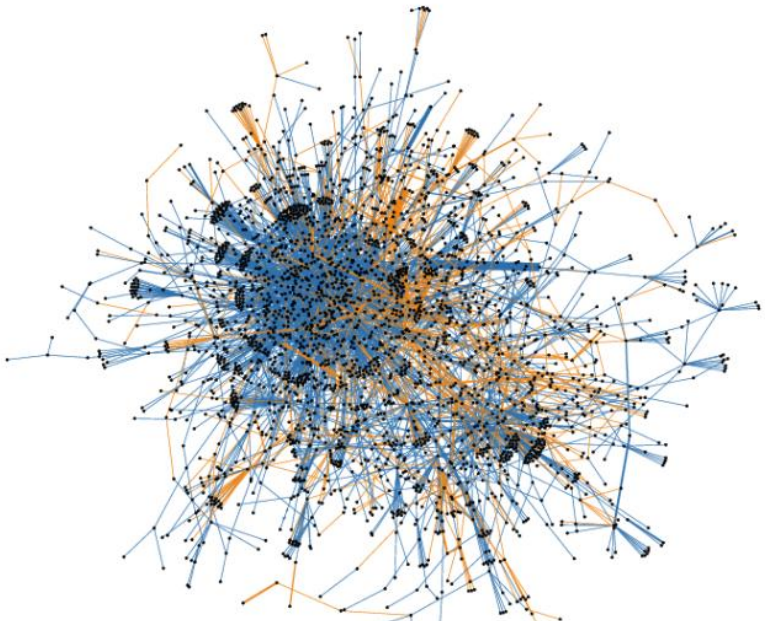


Рисунок 15. Распространение сообщений с тэгом #teaparty в Twitter [**106**] с помощью Truthy

3.3.2 MentionMapp

MentionMapp – сделанная компанией Asterisq [**68**] на flash визуализация последних упомнинаний пользователей и тэгов в потоке Twitter-сообщений, при этом число упоминаний влияет на силу связи.

Из-за простоты запроса и результата, объект исследования может быть задан любым посетителем, к сожалению хронологической и структурной глубины настроить нельзя (Рисунок 17)

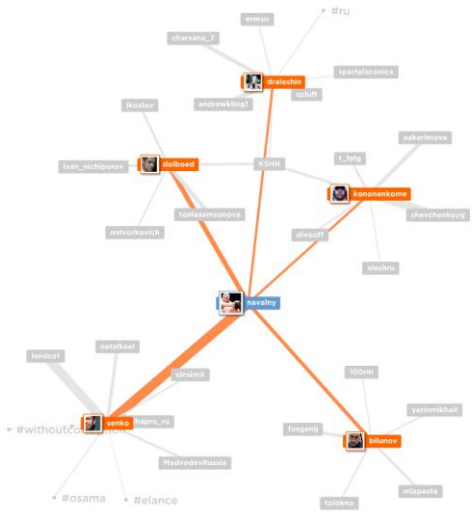


Рисунок 16. Визуализация упоминаний антикоррупционного активиста @navalny [**107**]

3.3.3 SocialCollider

Social Collider — визуализация Twitter-сообщений на Javascript под Google Chrome в виде горизонтально расположенных колонок, связанных с объектом исследования сущностей (последние упоминания, хэштэги и ссылки), которые заполняются точками, означающими сообщения в хронологическом порядке [**69**]

Сообщения связываются между собой по кривой, причём у каждой колонки она отмечена своим цветом. В случае взаимосвязи сообщения из одной колонки с другой, цитирование выделяется резкими прыжками кривой. К сожалению, разбиение на колонки не интуитивно и не показывает социальной структуры, а фокусируется лишь на частоте публикаций и возникающих взаимосвязях.

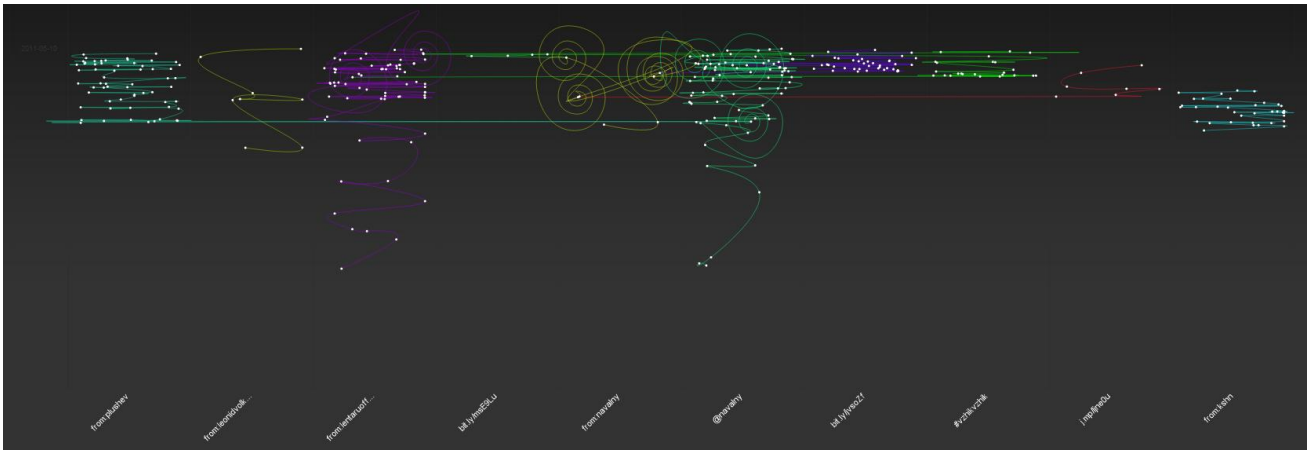


Рисунок 17. Сообщения близких к @navalny пользователей и их цитирование [**108**]

4. Практическая визуализация

Как было описано выше, сети обширны как по своим типам, так и по свойствам. Существующие приложения очень богаты функциональными возможностями, однако библиотек способных в режиме реального времени работать напрямую в браузере, крайне мало. Нам известна только библиотека Arbor.js [**70**], которая хотя и имеет хорошую анимацию, не имеет функциональности по анализу структуры графа.

4.1 Модель

В отличие от описанных выше инструментов мы предлагаем визуализировать не только статичную структуру социальной сети на основе друзей, или распространение одного каскада в сети как дерева, а рассматривать оба процесса одновременно как взаимно влияющие.

Пример эволюции ложного каскада описан [**71**]. Изначальное сообщение автора с цитатой МЛК [16] в результате превратилось в ложную цитату. Переход возник из-за желания опустить кавычки и сократить сообщение. Из-за доверия друзьям и сложности проверки истинности, цитата распространилась среди миллионов читателей менее чем за два дня. Визуализация такого процесса и интересует нас.

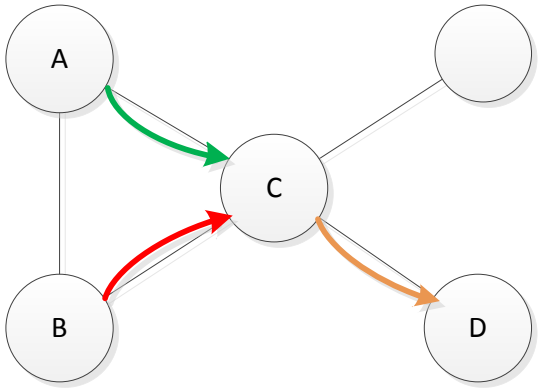


Рисунок 18. Концепт визуализации слияния двух каскадов (A↔C и B↔C) в новый C↔D поверх существующих связей. Слияние цвета подчёркивает объединение тем.

[16] Мартина Лютер Кинг младший, американский проповедник и активист за гражданские права, Нобелевский лауреат премии мира 1964 г.

Визуально мы предлагаем эволюционирующий каскад представлять как структурно динамичный граф сообщений поверх графа социальных связей, при этом частоту сообщений можно использовать в качестве веса нижестоящей связи

4.1.1 Важность микросил

Кроме решения визуализации мы предполагаем, что социальное поведение стимулирует поддержание и создание связей, которые характеризуют форму каскадов.

Эволюционирующий каскад должен оказывать **обратную связь** на вершины, которая стимулирует его повторное распространение или наоборот повышенную диффузию и даже сопротивление.

Упомянутая выше модель пространственной сегрегации [**37**] а также влияние согласованного смешения, усиленного предпочтительной связью [**72**] вкупе с влиянием каскада на формирование новых связей таким образом формирует сеть в макро масштабах.

Следовательно, для поддержания здоровой сети, политики [**73**] и управляющие компаниями [**74**] должны учитывать создание соответсвующей атмосферы.

4.2 Архитектура

Вычисление в браузере неэффективно при больших объёмах, тем не менее, для многоплатформенности и изучения небольших сетей, предлагаемое решение с HTML5 canvas и javascript более мобильно и подходит для визуализации объёмов в пределах 1000 вершин. Создание более серъёзной платформы описывается в работах на будущее.

В рамках этой работы реализуется приложение, которое занимается четырьмя функциями — добавлением вершин и рёбер, расчётом и перемещением их в двумерной плоскости, интерактивным управлением и расчётом свойств сети.

4.3 Структура данных

Добавления вершин и рёбер сделано двумя методами, которые могут вызываться как сразу при загрузке страницы, так и постепенно с асинхронной загрузкой или генератором случайного графа

Рисование напрямую связанное с индексацией (crawling) так же возможно, однако практически этот процесс неэффективен, поскольку большинство социальных сетей имеют временные ограничения[17]

Вершины хранятся в виде одномерного множества (javascript объект *nodes*) в котором по уникальному ключу вершины можно получить её объект типа *Node*, который детально реализует внутренние свойства и методы движения.

Рёбра хранятся в прямой и обратной матрице смежности для быстрого поиска как по источнику, так и по назначению.

4.4 Алгоритм силового расположения вершин

Как было уже описано [**52**], итеративные модели легко реализуется, но самый тривиальный имеет сложность $|V|^2$ из-за отталкивания вершины от всех вершин. Мы следуем описываемым решениям и используем алгоритм [**46**] для расчёта сил между двумя вершинами, однако суммарную силу, действующую на вершину, мы вычисляем рекурсивно до определённой глубины.

Максимальная глубина — диаметр подграфа, практически же достаточно использовать радиус, чтобы избегать схлопывания полых графов. Однако учитывая ранее показанную закономерность стремления эффективных диаметров социальных сетей к размеру 6, в данной работе использовалась глубина равная 2-3, в зависимости от размеров выборки.

[17] например Twitter API ограничен 150/350 запросами в час [**87**]

```
    function recursiveRepulsiveForceFlow(ForceSum, realNode, node, deepness,
level_multiplier, traversed_nodes){
            for(z in connections[node.ID]){
                dstNode=nodes[z];

                if(dstNode!=null && realNode.ID!=dstNode.ID

                    RepulsionVector = Vector(
                            - (dstNode.x-realNode.x),
                            - (dstNode.y-realNode.y)
                            );

                    AbsStep = fRepulsionCoefficient * iOptimalDistance *
iOptimalDistance / RepulsionVector.length();
                    if(AbsStep>max_movement_speed) {
                        AbsStep = max_movement_speed;
                    }

                    RepulsionVector.normalize();
                    RepulsionVector.multiply(level_multiplier);

                    RepulsionVector.multiply(AbsStep );

                    ForceSum.add(RepulsionVector);

                    deepness--;
                    if(deepness>0 && typeof( traversed_nodes[dstNode3.ID]
)=='undefined'){
                        traversed_nodes[dstNode3.ID]=1;
                        recursiveRepulsiveForceFlow( ForceSum, realNode,
dstNode, deepness, level_multiplier*level_multiplier, traversed_nodes);
                    }
                }
            }
    }

    recursiveRepulsiveForceFlow(ForceSum,node,node,2, 0.3, {});
```

Алгоритм 1. Рекурсивное вычисление суммарной силы отталкивания (ForceSum) для вершины (realNode), учитывая глубину рекурсии (deepness) , убывающий коэффициент влияния (level_multiplier) и пройденный маршрут (traversed_nodes).

4.4.1 Адаптивное трение

Энергетическая замкнутость с изначальным случайным позиционированием вершин — очевидная проблема, поскольку начальные позиции вершин получают случайную потенциальную энергию, которая благодаря действующим силам переходит в кинетическую энергию, в результате граф постоянно остаётся в движении поскольку идеальная частица по инерции пролетает точку стабильности.

Для этого вводится пошаговая потеря скорости (изменяемый параметр, по умолчанию 10%). Для того что-бы не скатиться в локальный минимум, изменяется коэффициент потери на прирост, тем самым создаётся колебательный процесс стабилизации всей системы. Для затухания используется обратная логарифмическая функция от общего шага системы, тем самым предполагается «температурная смерть вселенной». Шаг при котором потеря энергии будет полной есть корень уравнения

$$E = 1/log(\tau) - k$$

E – угасающий коэффициент прироста энергии
τ – шаг вычисления
k – коэффициент сдвига

Легко видеть, что при k=0.1 и E=0 → $\tau \approx 22027$

```
energyLoss=0.9;
if(node.acceleration.length()<0.1){
    node.energyGainStep++;
    if(node.energyGainStep >5){
        energyLoss = (1/Math.log(graph.step) - 0.1) / energyLoss;
        if(energyLoss<0) energyLoss=0;
    }
}
else{
    node.energyGainStep=0;
}
```

Алгоритм 2. Угасающий колебательный процесс потери энергии вершины во времени

4.5 Рисование и интерактивное взаимодействие

4.5.1 Масштабирование

Наблюдаемая Вселенная доступна от 10^{-35} до 10^{24} метров с различной степенью кластеризации материи на разных масштабах. Мы предполагаем, что структура социальных сетей организована с похожей сложностью [18].

[18] Так банкротство крупного предприятия может быть мало заметным при атомных масштабах и выделятся на более высоких, подобно взрыву сверхновых

Понятно, что при таких масштабах, подобную структуру невозможно визуализировать с линейным масштабированием.

Для масштабного рисования графа, с учётом фиксированного размера реального холста, вносятся глобальные переменные, имитирующие *виртуальное окно* просмотра:

- Координаты верхнего левого угла
- Коэффициент увеличения

Теперь при дискретном сигнале изменении масштаба (при кручении колёсика мышки с привязанным событием onmousewheel) необходимо поменять коэффициент увеличения и сдвинуть начальное положение, что бы масштабирование происходило во все четыре стороны. При этом для удобства пользователя, при увеличении масштаба, используется приспособляющееся масштабирование — пропорции сдвига окна зависят от положения курсора так, что после увеличения, координаты точка под курсором на экране осталась там же.

Легко видеть, что составив уравнение пропорциональности положения точки в старой и новой системе координат, можно найти сдвиг окна:

$$\frac{x_c}{w} = \frac{x_c + \Delta_x}{z \times w} \qquad \rightarrow \qquad \Delta_x = x_c(z-1)$$

Δ_x—сдвиг начальных координат окна,
x_c— расстояние от начала окна до курсора,
w—ширина окна,
z — шаг дискретного увеличения

Теперь достаточно добавить изменение положения и увеличения окна и при рисовании делать трансформацию координат вершин и связей в соответствии с положением окна. Шаг z мы выбрали равным 1.1 для плавного степенного масштабирования (легко видеть что при 10 шагах, ширина окна станет $w \times 1.1^{10}$)

4.5.2 Передвижение

Благодаря описанному выше использованию виртуального окна для масштабирования, стало возможным и передвижение по экрану. Для этого обрабатывается нажатие и отпускание клавиши мышки (onmousedown и onmouseup) с регистрацией временного состояния и где нажатие на экране произвелось. При передвижении мыши (событие onmousemove) сдвигается виртуальное окно на расстояние от изначального нажатия, с поправкой на коэффициент увеличения.

4.5.3 Выделение и динамические настройки

Выделение вершины играет важную роль при локальном анализе графа. Для выделения была дополнена обработка события onmouseup случаем, когда движения мышки не было, иначе при любом передвижении происходило бы выделение вершины. Фактически были взяты координаты курсора и в цикле по всем вершинам находится самая близкая, с сохранением её в массив выбранных вершин. Далее в процессе рисования идёт проверка, выбрана ли вершина с выделением её особым цветом.

Для более гибкой настройки рисования, была добавлена возможность управления некоторыми переменными с помощью бегунков в UI. Можно настроить процент потери кинетической энергии за шаг, коэффициент отталкивания, размер вершины и видимость самих вершин, рёбер и подписей. Кроме того для анализа ядра сети, мы добавили фильтр вершин по их степени.

4.6 Характеристики

Для того что-бы понимать внутреннюю структуру, был добавлен элементарные показатели — число вершин и связей и производная от них — плотность. Суммарная кинетическая энергия показывает общую динамику графа к стабилизации и вычисляется периодически, отдельно от цикла рисования.

График распределения степеней также динамически рисуется с помощью google.visualization.DataTable [**75**], матрицу смежности – с помощью canvas. Алгоритм

вычисления коэффициента используется более пригодный для комбинаторного вычисления [**76**]

```
function getClusteringLevel (){
    Csum=0;
        for(i in nodes){
            rank = getNodeRank(nodes[i].ID);
            if(rank>1){
                Csum = Csum + getNodeMutualFriendCount(nodes[i].ID)/(rank *
(rank-1));
            }
        }
        return Csum/count(nodes);
    }

function getNodeMutualFriendCount (nodeID){
     sum=0;
        for(source in connections[nodeID]){
            for(target in connections[source]){
                if(typeof(connections[nodeID][target]) !='undefined'){
                    sum = sum + 1;
                }
            }
        }
        return sum;
    }
```

Алгоритм 3. Вычисление степени кластеризации

5. Результаты

5.1 Twitter

Twitter – международная социальная сеть принадлежащая американской компании. Из-за открытости данных, популярности, доступного API и простого хронологического списка сообщений, была выбрана социальная сеть Twitter для анализа как структуры друзей, так и распространения сообщений.

5.1.1 Друзья

Для анализа небольшой выборки данной сети, был написан рекурсивный сборщик данных пользователей и их друзей на javascript с фильтрацией местоположения пользователя по Эстонии. После мы разделили выборку и отображение, добавив сохранение через JSON в локальную базу данных MySQL с чтением из неё для рисования.

Однако поскольку Twitter на момент работы не позволял напрямую фильтровать пользователей по местоположению, то пришлось рекурсивно спрашивать друзей и у каждого друга – местоположение.

Дойдя до 923 просмотренных пользователей и 30539 в ожидании, мы прекратили выборку за неэффективностью из-за ограничения на скорость Twitter API. У полученного графа 218 пользователей и 1657 связей и коэффициент кластеризации 0.35.

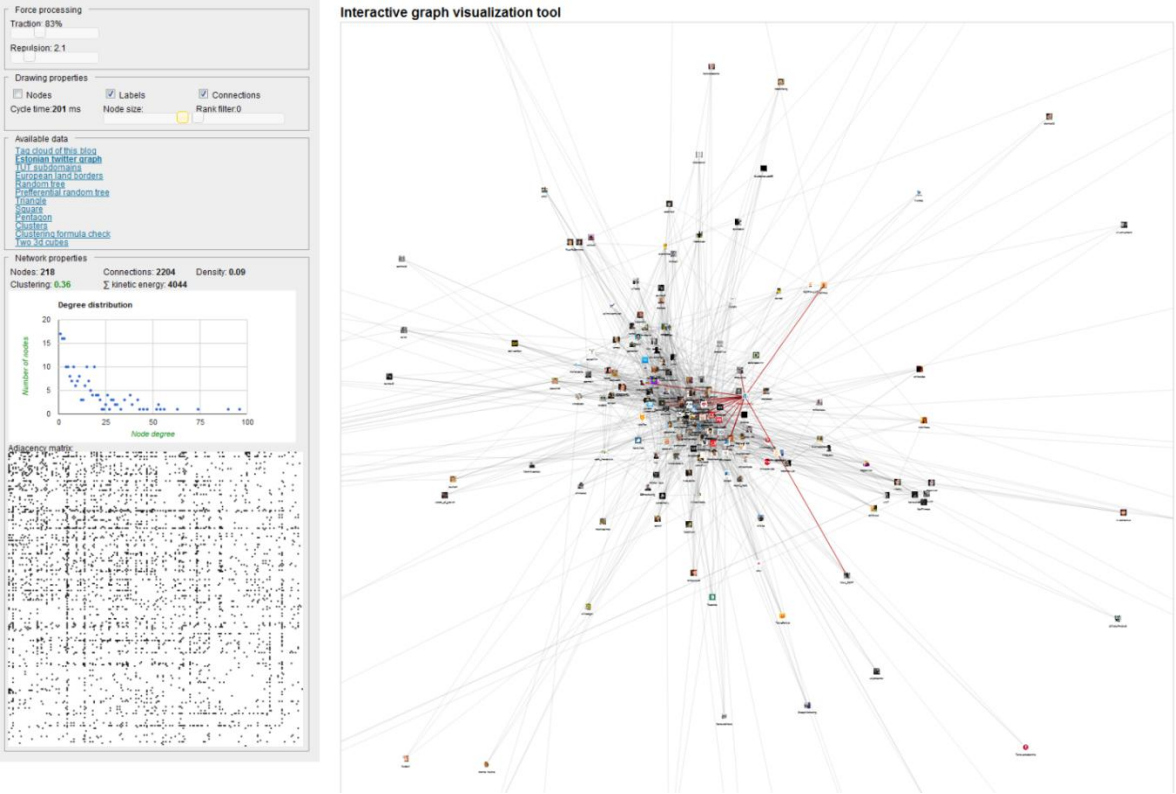


Рисунок 19. Внешний вид интерфейса инструмента для рисования и анализа. Слева панель управления. На графе 218 эстонских twitter-пользователей. Красным обозначен выделенный пользователь и его связи

5.1.2 Сообщения

Поскольку в предыдущем случае регистрация новых сообщений потребовала бы большей сложности (мониторинг каждого пользователя), была выбрана регистрация потока всех сообщений, где источник находится в Таллинне.

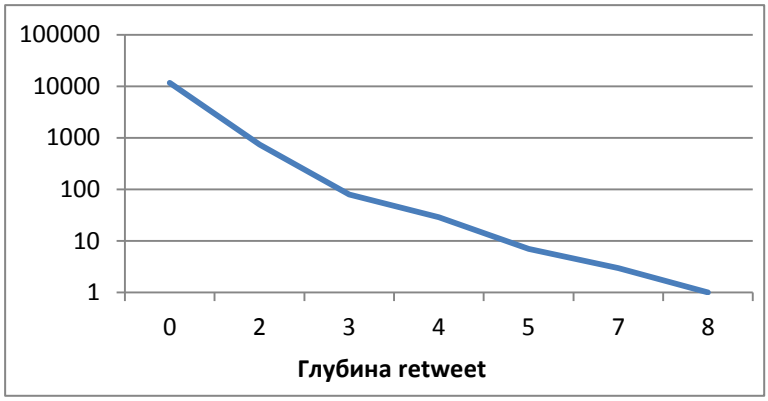


График 5. Логарифмическая зависимость глубины retweet от числа сообщений в Таллинне 5-22 мая 2011 г.

За промежуток 5-22 мая, благодаря RSS-потоку Twitter-поиска, были зарегистрированы 12643 сообщений, из них 6.89% — retweet и 22.56% — направленные сообщения, причём глубина *retweet* имеет логарифмическую зависимость.

В результате визуализации было получено множество независимых графов (степень кластеризации 0.02) с практически отсутствующим GCC. Это может объяснятся как слишком малым временем для анализа связей на основе сообщений, так и децентрализованностью писавших в Таллинне пользователей (туристы, ошибки API со включением внешних пользователей в поток и тому подобное)

Приведённая сеть retweet-сообщений (Рисунок 21) лишь часть полноценного каскада, поскольку пользователи могут развивать тему другими способами.

К сожалению даже доступная публичная переписка вынуждает использовать методики классификации, которые в данной работе не рассматриваются. В перспективе определение близости сообщений позволит связать их в единую временную цепочку каскада.

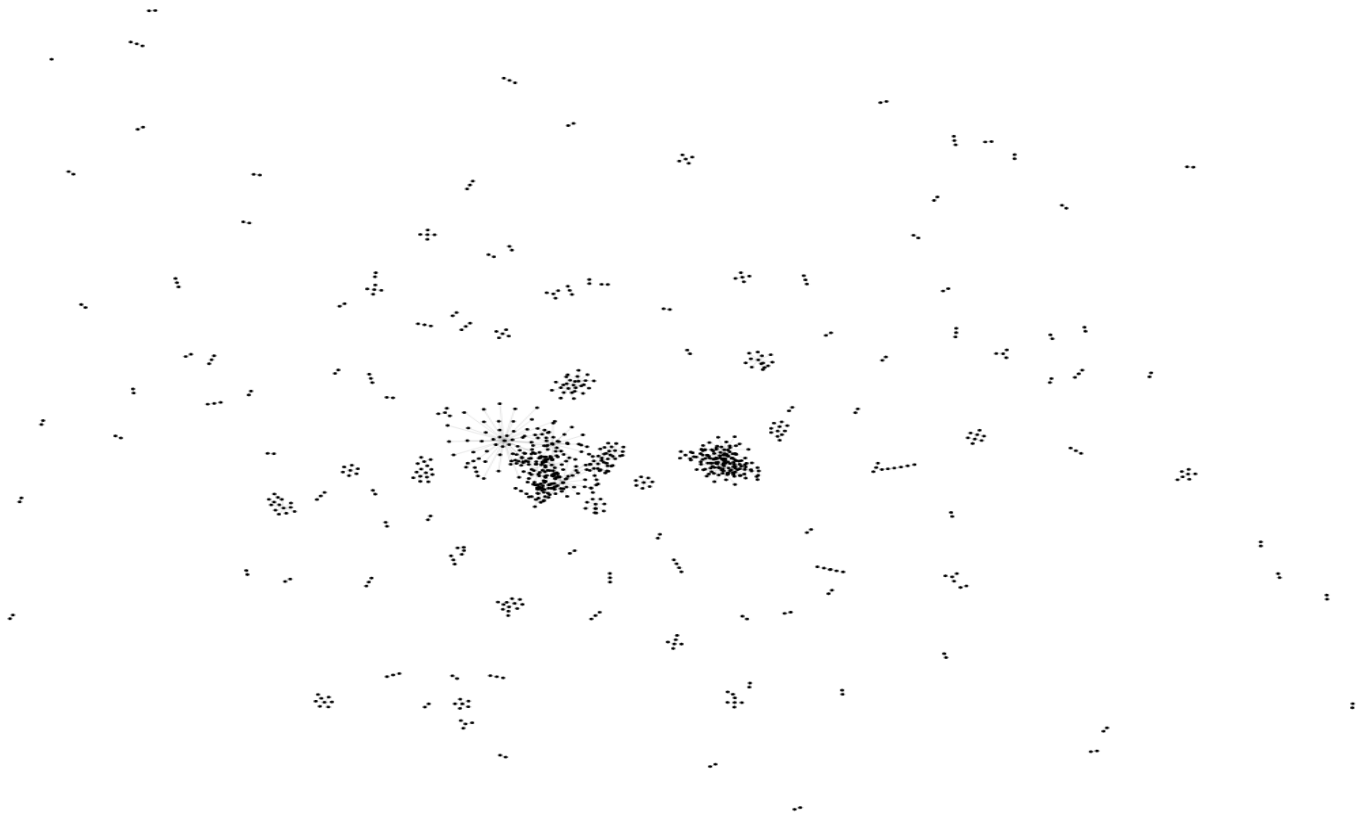


Рисунок 20. Несцепленная сеть 880 пользователей и 1044 retweet-сообщений в Таллинне в промежутке 5-22 мая 2011

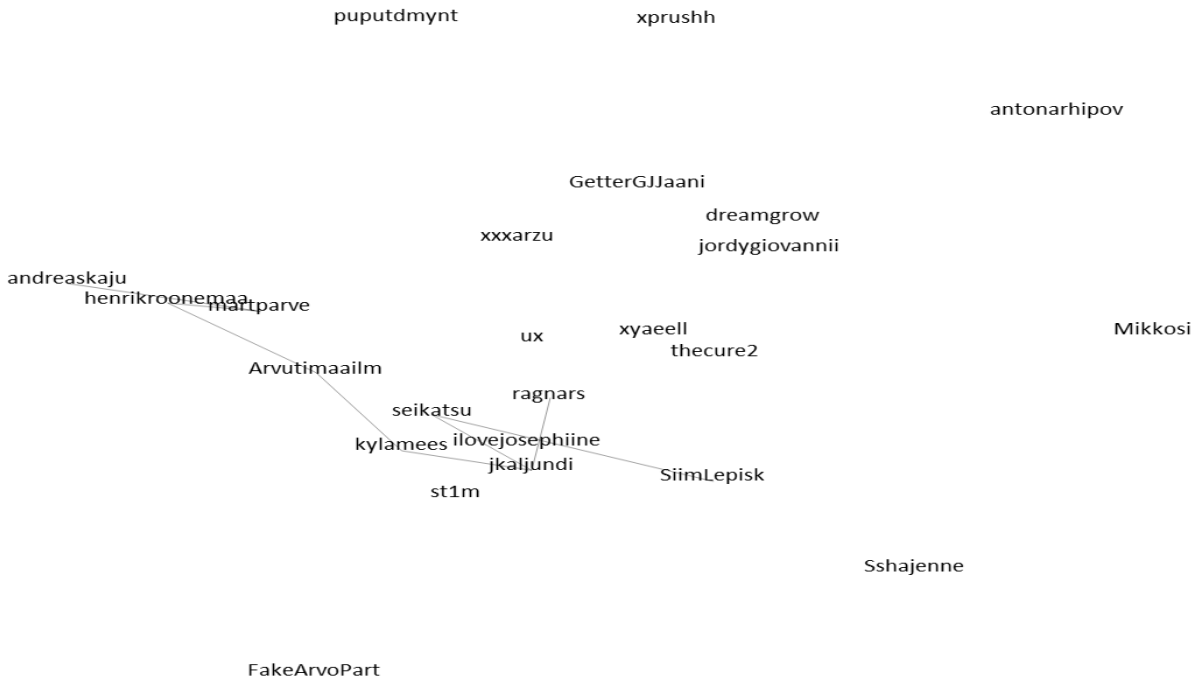


Рисунок 21. Наиболее активные retweet-пользователи Таллинна (степень более 8)

5.2 Анализ Pling.ee

Pling.ee — эстонская молодёжная социальная сеть, принадлежащая телекоммуникационной компании Elisa Eesti AS. Сайт позволяет добавлять текстовые, фото- и видео-сообщения через телефон (SMS и MMS), определять своё местоположение а также читать друзей и получать оповещения. Кроме того есть фильтры сообщений, настройки приватности, сообщества.

Рисунок 22 . Внешний вид интерфейса просмотра сообщений

Данная работа проведена с согласия правления Elisa Eesti AS и не содержит личных данных пользователей которые нарушали бы их приватность. Из-за больших объёмов, было решено отказаться от использования описанного выше инструмента визуализации и использовать Gephi

5.2.1 Пользователи

На момент анализа, база данных состояла из 136375 аккаунтов и 437018 связей, однако если учитывать только пользователей (без сообществ) и только тех кто хотя бы одиножды зашли на сайт, то размер уменьшится до 78458 вершин.

Географически расположение пользователей проводилось с помощью экспорта координат в csv формат и их визуализации с помощью специального сайта обработки массивных данных [**77**]

Распределение напрямую связано с урбанизацией и незначительно присутствует в Финляндии, однако интересно отметить, что некоторые области практически лишены покрытия — Маарду, район Нымме в Таллинне, южное Кохтла-Ярве.

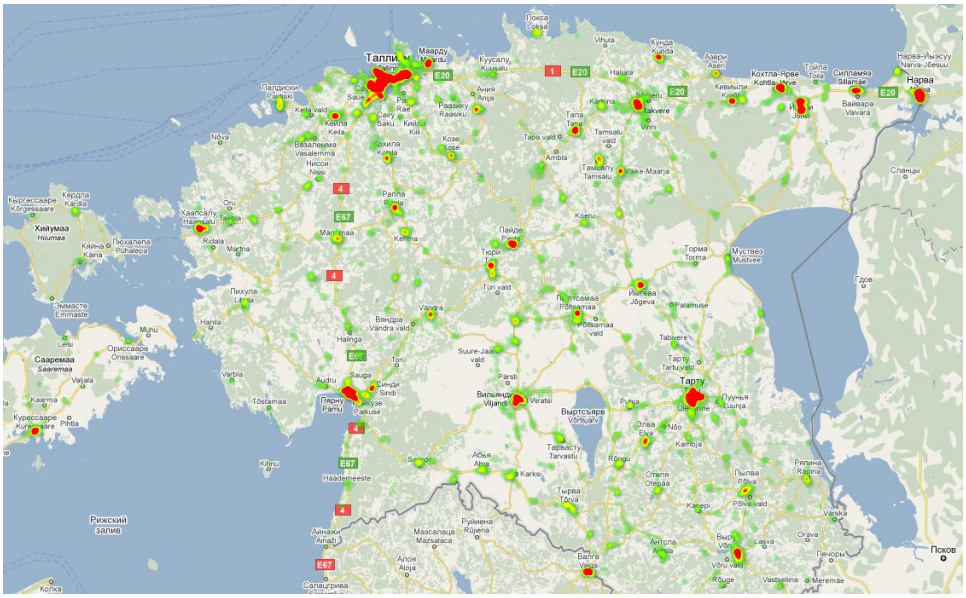


Рисунок 24. Heatmap распределения пользователей в Эстонии

Рисунок 23. Точечное географическое распределение пользователей в том числе в Финляндии

Из 104212 пользователей указавших свой пол, 54.19 % мужчины. Интересна и зависимость числа пользователей от указанного ими возраста, хотя очевидный максимум должен быть в районе 19-21 лет, есть и пики в 41 и в 1 лет соответственно. Если возраст в 1 год можно понять первым

значением в UI элементе, то возраст в 41 год, по всей видимости, объясняется повреждением данных в ранней стадии разработки, когда сохранение даты было в unix timestamp системе и при конвертировании стало 1970 годом – началом системы исчисления.

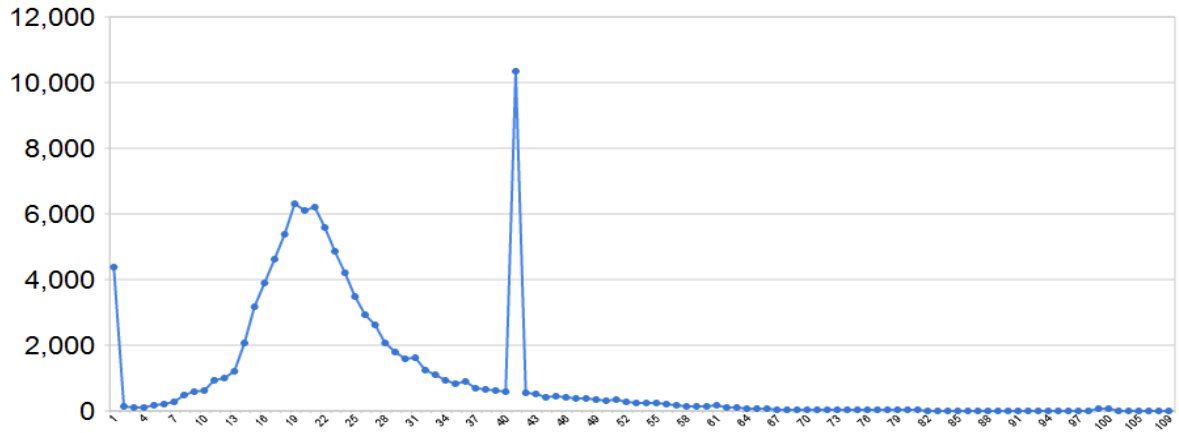


Рисунок 25. Распределение числа пользователей в зависимости от указанного ими возраста

5.2.2 Друзья

Структура на основе заявленных друзей выглядит естественно для здоровой социальной сети — выборка 74809 вершин образует плотный GCC. Из-за многочисленных хабов и большого числа переплетений, сложно заметить детали. Максимальная степень вершины равна 4767.

Свойства	Значение
Средний коэффициент кластеризации	0.135
Средняя степень вершины	4.313
Диаметр GCC	20
Средний диаметр GCC	5.38

Таблица 2. Свойства сети на основе данных о друзьях pling.ee

Была проведена более агрессивная выборка наиболее активных пользователей (телефон и следовательно пользователь активен в течение двух недель) из 46259 вершин и 99764

связей. На основе результатов (Таблица 2, Рисунок 26) можно сказать, что наблюдается *small world effect*, а степень кластеризации близка к сети рассылки почты [**12**].

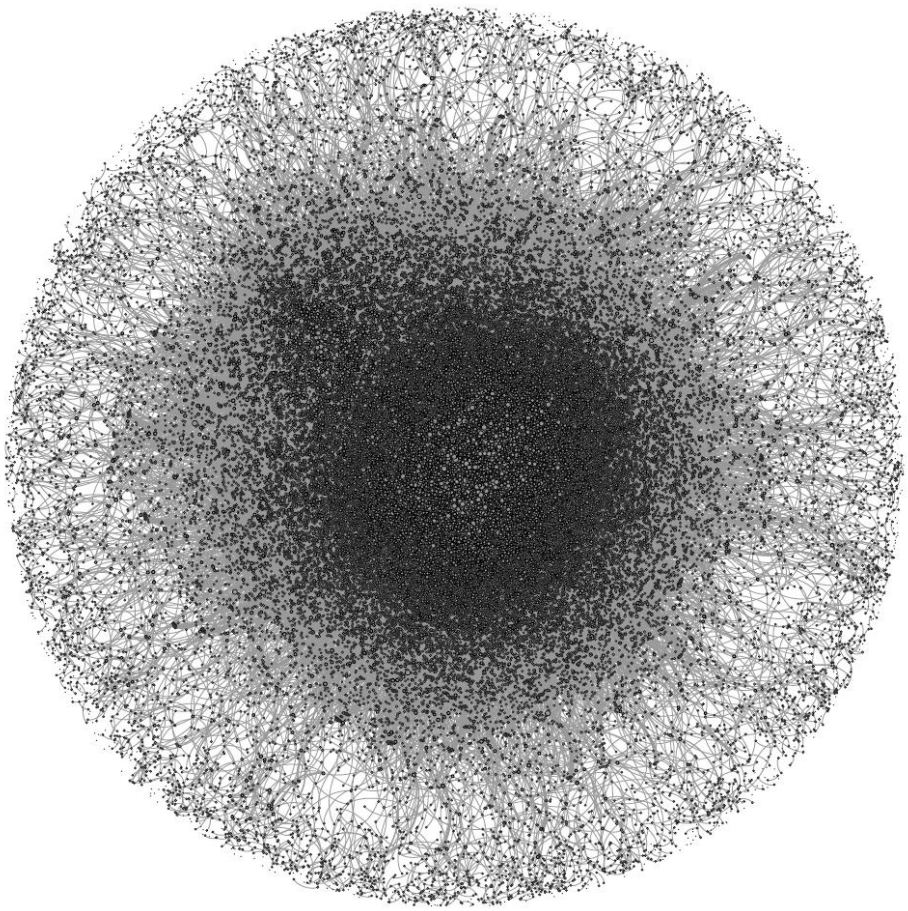


Рисунок 26. Сеть друзей Pling.ee (75 тыс. вершин) на момент 17.05.2011. Сделано с помощью Gephi с алгоритмом Yifan Hu

5.2.3 Сообщения

За всё время работы база данных состоит из 18,6 млн сообщений. Из них 893 тыс было создано с 1 мая 2011 00:00 до 18 мая 2011 12:50, из них 790 тыс — приватные, т.е. 89%.

График 6. Публичные сообщения по структуре

Таким образом, в день создаётся около 6 тысяч **публичных** сообщений, из них 26% на кириллице (транслит не подсчитывался) и только 14,3% из веб-браузера. Тематически практически все сообщения касаются бытовых отношений.

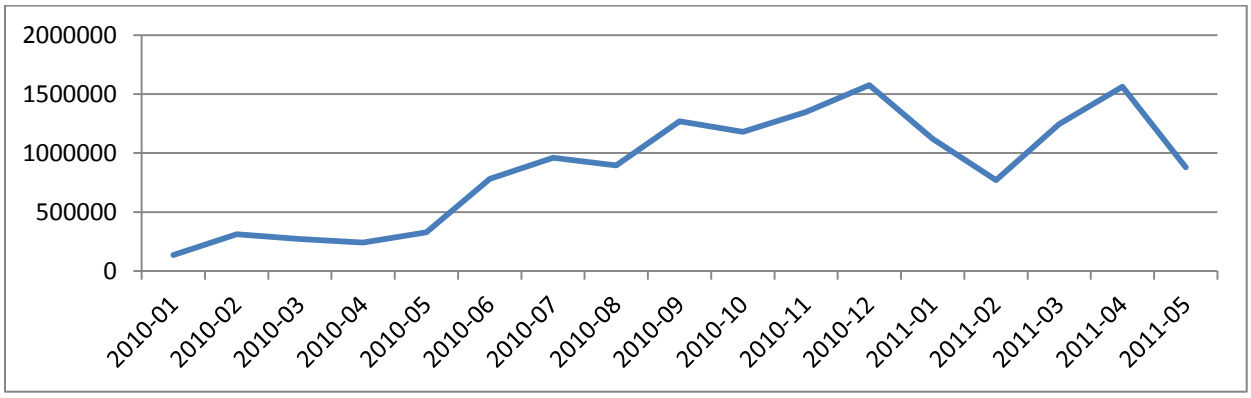


График 7. Количество сообщений в зависимости от времени

Понятно, что не смотря на здоровые структурные показатели сети, тематика романтического содержания и интерфейс общения (телефон), препятствуют возникновению каскадов и рассматривать систему как платформу для микроблога нельзя, несмотря на наличие функциональности сообществ, тэгов и изображений.

Параметр	Значение
Средний коэффициент кластеризации	0.043
Средняя степень вершины	2.202
Диаметр GCC	38
Средний диаметр GCC	13.009

Таблица 3. Свойства сети на основе публичных сообщений pling.ee за 18 дней

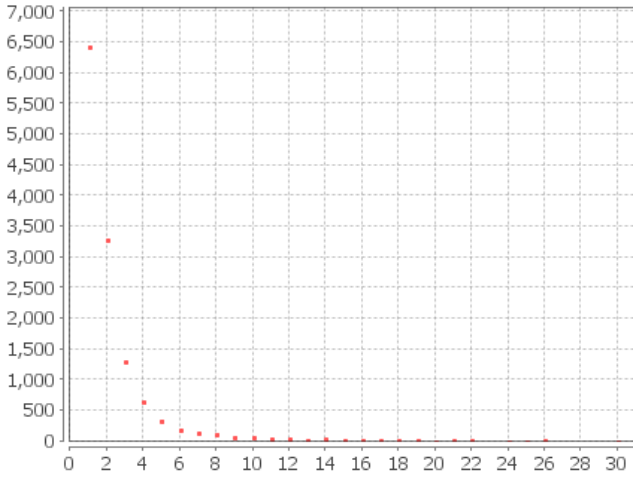


График 8. Распределение степеней графа связей на основе публичных сообщений pling.ee за 18 дней

Для сравнения декларируемых социальных связей с реальными и нахождения более закономерной структуры, мы сделали выборку 84245 публичных направленных сообщений с 1 по 18 мая 2011 г., при этом мы получили 12686 участников.

Из результатов (Таблица 3) чего можно сделать предсказуемый вывод, что общие показатели сети на основе сообщений слабее (меньше кластеризация, больше диаметр), чем сеть друзей, более интересно то, что степень кластеризации более близка к электрическим сетям [**12**]. Также бросается в глаза согласованное смешение по языку — наглядно видны два кластера (Рисунок 27)

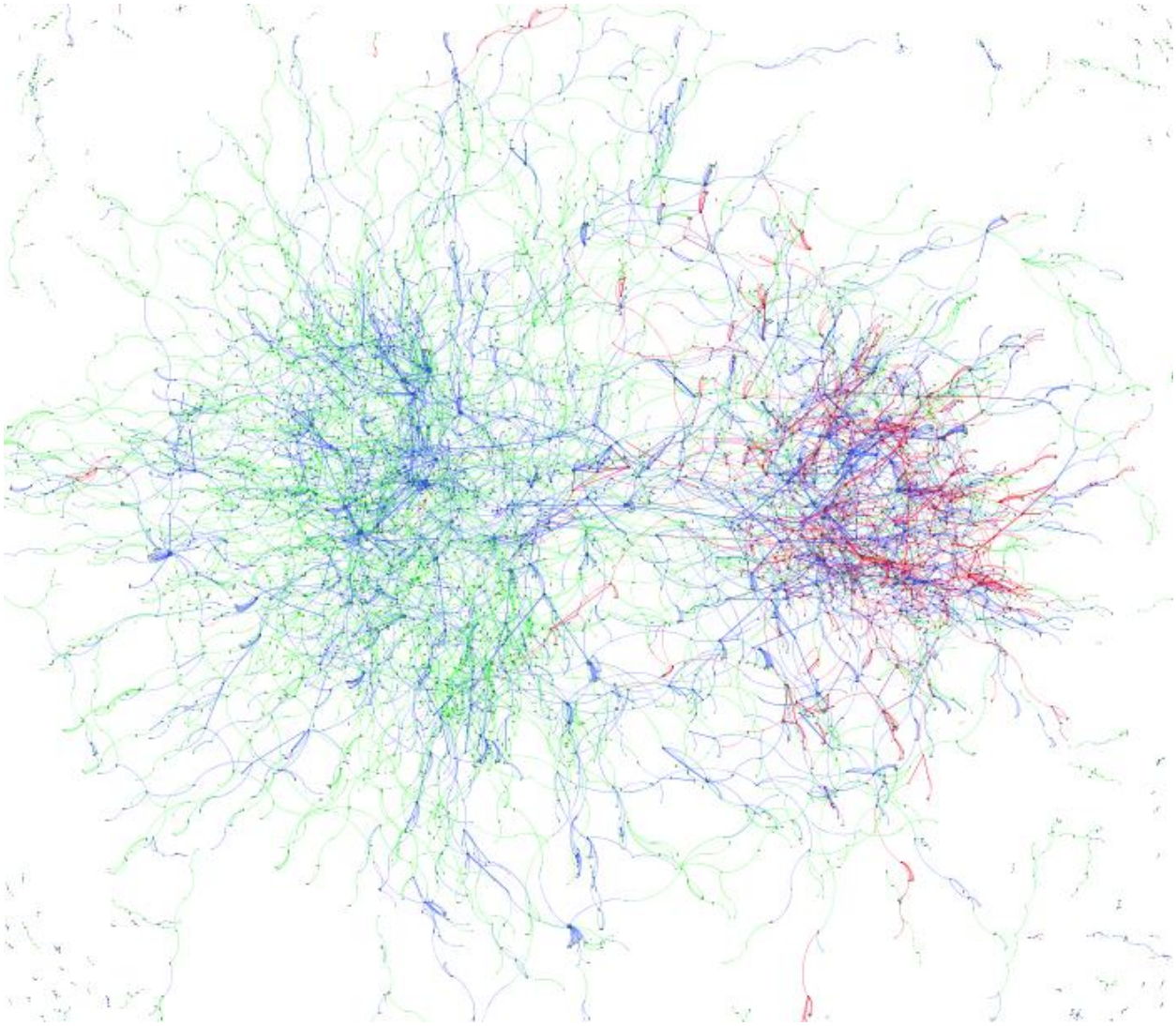


Рисунок 27. GCC графа из 12 тысяч участников публичных сообщений pling.ee за 18 дней, красным цветом обозначены пользователи больше использующие кириллицу (4%), синим - специальные символы эстонского алфавита (20%), зелёным – все остальные

5.3 Дальнейшие работы

Для детального анализа распространения каскадов необходима более широкая выборка данных, которая учитывала бы связи на протяжении больших промежутков времени.

Для масштабирования публичных или очень объёмных приложений визуализации, надо использовать серверные вычисления и отображать клиенту статичную картинку, например через неблокирующий сервер nodejs [**78**] написанный на C++ и Python, позволяющий запускать серверный javascript и способный работать с canvas с помощью расширения [**79**].

Серверные вычисления позволили бы ускорить вычисление движение огромных графов в реальном времени через параллельные потоки и прямой доступ к использованию видеокарт через OpenGL или CUDA, дали бы возможность использовать графовые базы данных [**80**], инструменты как Gephi.

Кроме масштабирования объёмов, необходимо многомерно анализировать уровень приложений в рамках исследуемой социальной сети, т.е. IM, Email, физический контакт одновременно для более чёткой картины.

Уже сейчас некоторые социальные сети разделяют тип социального поведения от простых действий Reshare и Like до выражения благодарности, согласия [**81**] и оценки [**82**]. Нельзя не забывать и о свойствах предпочтительного связывания. Влияние таких микро-действий в макромасштабах сообщества также хорошая область исследования.

Наиболее интересной представляется область распространение каскадов с классификацией и оценкой близости темы сообщений и учитыванием тематической авторитетности источников.

Кроме социальных сетей, была испопробована визуализация информационных сетей, выборка которых была получена с помощью индексирования страниц [**83**]. Сравнение

потоков информации или транспорта было бы любопытно для выявления общих закономерностей в перколяционной теории сетей.

В данной работе также была только поверхностно затронута кластеризация, поэтому в дополнение к фильтрации по степени вершины, для детального анализа сети, необходим и фильтр по дендрограмме кластеров.

6. Вывод

Рассмотрев свойства и особенности социальных сетей, мы предложили свой алгоритм и инструмент для рисования и навигации. Проделав извлечение данных двух социальных сетей с фокусом на Эстонию, была показана большая наглядность структуры сети при использовании динамической сети общения вместо сети друзей.

Сравнив Twitter и Pling, были замечены кардинальные отличия в существовании каскадов. Тем не менее в обоих случаях наблюдался эффект согласованного смешения (*assortative mixing*) по языку, степенной закон распределения степеней вершин (*degree distribution*) и малый диаметр социального графа (*small world effect*)

В заключение, мы считаем что возникновение каскадов зависит не от топологии, а от аудитории (источника), интерфейса (проводника) и самое главное - силы сообщения. Для визуализации эволюции каскадов необходимо применять тематическую классификацию сообщений.

Summary

Any visualization requires knowledge of underlying field. In this thesis, social networks are described, along with its analysis using graph theory. A special attention is given to importance of interdependent-message spread over time (cascades).

The primary goal is to introduce primary characteristics that influence structure and dynamics of network, and existing applications that use them, and build own instrument as a result.

After describing more than 10 similar tools, thesis introduces browser-based tool for visualization and analysis of networks that uses HTML5 canvas. Main structure, drawing- and interaction- algorithms are described.

Comparison is described using retrieved two data sets of Estonian Twitter users — based on friend networks and on message flow with Tallinn as location (using existing API). These data sets are visualized using created tool and consist of less than 1 thousand nodes.

For broader network analysis of estonian Pling.ee social network, an opensource application Gephi was used. Detailed analysis of 70 thousand users in friend network and a 12 thousand users in message network is presented.

Although second dataset had bigger size, no cascades were found, even though many characteristics existed (in particular language assortative mixing)

Because of selected technology, created instrument is very limited in computation and memory size, but is much more mobile than Gephi. Evolutionary cascade visualization is presented as concept using text classification.

Keywords: graph theory, social network, visualization, clustering, classification

Resümee

Iga visualiseering eeldab teadmisi antud valdkonnas. Selles töös kirjeldatakse sotsiaalvõrgustikke koos graafiteooriat kasutava analüüsiga. Erilist tähelepanu pööratakse omavahel sõltuvate sõnumite levikule ajas (kaskaadile).

Põhieesmärk on tutvustada põhilisi näitajaid, mis mõjutavad võrgu struktuuri ja dünaamikat, ning olemasolevaid rakendusi, ja lõpptulemusena luua oma tööriist.

Peale rohkem kui 10 sarnase vahendi kirjeldamist tutvustatakse veebipõhist tööriista, mis oskab visualiseerida ja analüüsida võrgustikku, kasutades selleks HTML5-canvas elementi. Kirjeldatakse põhistruktuuri ning joonistus- ja interaktsiooni-algoritme.

Võrdluses tuuakse kaks Eesti Twitteri-kasutajate andmekogumit, mis põhinevad sõprade võrgustikel ja sõnumite voolus asukohaga Tallinnas. Andmekogumid visualiseeritakse loodud tööriistaga ja koosnevad vähem kui 1000 sõlmest.

Kohaliku Pling.ee sotsiaalvõrgustiku laiemaks analüüsiks kasutatakse vabavaralist programmi Gephi. Esitatud on 70 000 kasutajaga sõprade ja 12 000 kasutajaga sõnumite võrgustike analüüs.

Kuigi viimane andmekogum oli suurem, ühtegi kaskaadi ei leitud, vaatamata et paljud omadused olid esindatud (eriti keele-põhine segunemine).

Valminud tööriist on valitud tehnoloogia tõttu arvutusvõimsuses ja kasutatava mälu hulga poolest piiratud, aga mobiilsem kui Gephi. Evolutsioonilise kaskaadi visualiseering on esitatud kontseptsioonina kasutades teksti klassifitseerimist.

Võtmesõnad: graafiteooria, sotsiaal võrgustik, visualiseerimine, klastrite loomine, klassifitseerimine

Краткое изложение

Всякая визуализация подразумевает знание предметной области. В данной работе рассматриваются социальные сети и их анализ с помощью теории графов. Особое внимание ставится на важность распространения зависимых сообщений - каскадов.

Основная цель работы – ознакомится с главными свойствами, которые влияют на структуру сети и существующими приложениями, в результате построив свой инструмент.

Рассмотрев более 10 аналогов, описывается создание инструмента по визуализации и анализу сети в браузере с помощью HTML5 canvas. Описывается структура и некоторые алгоритмы рисования и интерактивного управления.

Для сравнения используются две выборки из подсети эстонских пользователей Twitter — на основе друзей и на основе потока сообщений Таллинна (благодаря существующим API). Эти выборки визуализировалась в браузере с помощью созданного инструмента и состояла из менее 1 тысячи вершин.

Для анализа более широкой выборки эстонской сети Pling.ee использовалась программа с открытым исходным кодом Gephi. Приведён детальный анализ 70 тысяч пользователей в общей сети друзей, а также сеть публичных сообщений 12 тысяч пользователей.

Не смотря на большие объёмы, в последней не было обнаружено каскадов, хотя и наблюдались многие характерные свойства (в частности согласованное смешение по языку).

Из-за выбранной технологии, представленный инструмент ограничен в мощностях и функциональности, однако он мобильнее Gephi. Кроме этого, предлагается концепт анализа эволюционирующих каскадов с помощью классификации текстов.

Ключевые слова: теория графов, социальная сеть, визуализация, кластеризация, классификация

Источники и литература

[1] Gregory Piatetsky-Shapiro, Usama Fayyad, and Padhraic Smyth, "Knowledge discovery and data mining: Towards a unifying framework," Menlo Park, CA, 1996.

[2] Matt Ridley, *The Origins of Virtue: Human Instincts and the Evolution of Cooperation*.: Penguin books, 1997.

[3] Tim Berners-Lee. Decentralized Information Group. [Online]. http://dig.csail.mit.edu/breadcrumbs/node/215

[4] Andre McKenzie et al., "Transmission Network Analysis to Complement Routine Tuberculosis Contact Investigations," *American Journal of Public Health*, October 2006.

[5] Nicholas Christakis and James Fowler, "The spread of obesity in a large social network over 32 years," *The New England Journal of Medicine*, pp. 370-379, 2007.

[6] James Fowler and Nicholas Christakis, "Dynamic spread of happiness in a large social network:longitudinal analysis over 20 years in the Framingham Heart Study," *British Medical Journal*, vol. 337, pp. 1-9, 2008.

[7] Lobbying Patterns in Healthcare Reform Around Key Senators. [Online]. http://www.orgnet.com/lobbying.html

[8] Matt Mohebbi et al. (2011, May) Google Blog. [Online]. http://correlate.googlelabs.com/whitepaper.pdf

[9] Stanley Milgram, "The small world problem," *Psychology Today*, 1967.

[10] Jure Leskovec and Eric Horvitz, "Planetary-Scale Views on an Instant-Messaging Network," 2007.

[11] Emden Gansner, Yifan Hu, and Steven Kobourov, *GMap: Drawing Graphs as Maps*.: AT&T Labs-Research, 2009, http://arxiv.org/abs/0907.2585.

[12] Mark Newman, "The structure and function of complex networks," 2003.

[13] Paul Erdős and Alfréd Rényi, "On the evolution of random graphs," 1960.

[14] E.N. Gilbert, "Random graphs," *Annals of Mathematical Statistics*, no. 30, pp. 1141-1144, 1959.

[15] Ray Solomonoff and Anatol Rapoport, "Connectivity of Random nets," *Bulletin of*

Mathematical Biology, vol. 13, no. 2, pp. 107-117, 1951.

[16] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos, "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations," 2005.

[17] Barabasi Albert-Laszlo Barabasi and Réka Albert, "Emergence of scaling in random networks," *Science*, pp. 509-512, 1999.

[18] Mary McGlohon, Leman Akoglu, and Christos Faloutsos, "Weighted Graphs and Disconnected Components," 2008.

[19] Steven Skiena, *The algorithm design manual*, 2nd ed.: Springer-Verlag, 2008.

[20] Edsger Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, pp. 269-271, 1959.

[21] Michael Fredman and Robert Tarjan, "Fibonacci Heaps And Their Uses In Improved Network Optimization Algorithms," in *25th Annual Symposium on Foundations of Computer Science*, 1987.

[22] Ravindra Ahuja, Kurt Mehlhorn, James Orlin, and Robert Tarjan, "Faster Algorithms for the Shortest Path Problem," *Journal of the Association for Computing Machinery*, vol. 37, no. 2, 1990.

[23] Robin Dunbar, "Neocortex size as a constraint on group size in primates," *Journal of Human Evolution*, vol. 22, no. 6, pp. 469-493, June 1992.

[24] Ulrik Brandes and Thomas Erlebach, *Network analysis:methodological foundations*.: Springer-Verlag, 2005.

[25] Duncan Watts and Steven Strogatz, "Collective dynamics of "small-world" networks," *Nature*, vol. 393, pp. 440-442, 1998.

[26] Vito Latora and Massimo Marchiori, "Efficient Behavior of Small-World Networks," *Physical review letters*, vol. 87, no. 19, November 2001.

[27] Wayne Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, pp. 452-473, 1977.

[28] Jan Fleischhauer. (2011, April) Nazi War Crimes as Described by German Soldiers. [Online]. http://www.spiegel.de/international/germany/0,1518,755385,00.html

[29] Julian Assange and Kristinn Hrafnsson. Collateral Murder. [Online]. http://www.collateralmurder.com/

[30] Kenneth Morrison, *Marx, Durkheim, Weber: Formations of Modern Social Thought: Foundations of Modern Social Thought*.: Sage publishing, 1995.

[31] Arthur Koestler, *The Act of Creation*.: Penguin Books, 1964.

[32] B. Ryan and N. Gross, "The Diffusion of Hybrid Seed Corn in Two Iowa Communities," *Rural Sociology*, pp. 15-24, 1943.

[33] Frank Bass, "A New Product Growth Model for Consumer Durables," *Management Science*, pp. 215-227, 1969.

[34] Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos, "Epidemic Spreading in a Real Networks: An Eigenvalue Viewpoint," 2003.

[35] James Coleman, "Social Capital in the Creation of Human Capital," *American Journal of Sociology*, 1988.

[36] Mark Granovetter, "The Strength of Weak Ties," *American Journal of Sociology*, pp. 1360-1380, 1973.

[37] Thomas Schelling, "Models of Segregation," *American Economic Review*, pp. 488-493, 1969.

[38] Eli Pariser, *The Filter Bubble: What the Internet is Hiding From You*.: Penguin Press, 2011.

[39] Thomas Malone, Robert Laubacher, and Chrysanthos Dellarocas, "Harnessing Crowds: Mapping the Genome of Collective Intelligence," 2009.

[40] James Surowiecki, *Wisdom of Crowds*.: Anchor Books, 2004.

[41] Michael Kaufmann and Dorothea Wagner, *Drawing Graphs: Methods and Models*.: Springer-Verlag, 2001.

[42] Kozo Sugiyama, Shojiro Tagawa, and Mitsuhiko Toda, "Methods for visual understanding of hierarchical systems," *IEEE transactions on Systems, Man and Cybernetics*, pp. 109-125, 1981.

[43] Peter Eades, "A heuristic for graph drawing," *Congressus Numerantium*, p. 42, 1984.

[44] Ignacio Alvare-Hamelin, Luca Dall'Asta, Allain Barrat, and Allesandro Vespignati, "k-core decomposition: a tool for the visualization of large scale networks," 2005.

[45] Giuseppe Di Battista, *Graph Drawing*.: Prentice, 1999.

[46] Thomas Fruchterman and Edward Reingold, "Graph drawing by Force-directed placement," *Software - practice and expirience*, pp. 1129-1164, 1991.

[47] William Tutte, "How to draw a graph," *London Math. Society*, 1962.

[48] Tomihisa Kamada and Satoru Kawai, "An algorithm for drawing general undirected graphs," *Information Processing Letters 31*, pp. 7-15, 1989.

[49] Kozo Sugiyama and Kazuo Misue, "Graph Drawing by the Magnetic Spring Model," *Journal of Visual Languages & Computing*, vol. 6, no. 3, pp. 217-231, 1995.

[50] Ron Davidson and David Harel, "Drawing Graphs Nicely Using Simulated Annealing," *ACM Transactions on Graphics*, vol. 15, no. 4, pp. 301-331, October 1996.

[51] Jürgen Branke, Frank Bucher, and Harmut Schmeck, "Using Genetic Algorithms for Drawing Undirected Graphs," 1996.

[52] Yifan Hu, "Efficient and High Quality Force-Directed Graph Drawing," *The Mathematica Journal*, vol. 10, pp. 37-71, 2005.

[53] James Moody. Social Networks: A statistical view. [Online]. http://www2.research.att.com/~volinsky/Graphs/slides/handcock1.pdf

[54] Aric Hagberg et al. NetworkX. [Online]. http://networkx.lanl.gov/

[55] Dimitris Kalamaras. Social Network Visualizer. [Online]. http://socnetv.sourceforge.net/

[56] Karen Stephenson and Marvin Zelen, "Rethinking centrality: Methods and examples," *Social Networks*, vol. 11, pp. 1-37, 1989.

[57] Jure Lescovec. Stanford Network Analysis Platform. [Online]. http://snap.stanford.edu/

[58] Jack Edmonds and Richard Karp, "Theoretical improvements in the algorithmic efficiency for network flow problems," *Journal of the ACM*, 1972.

[59] Ravindra Ahuja, Thomas Magnanti, and James Orlin, *Network Flows: Theory, Algorithms, and Applications*.: Prentice Hall, 1993.

[60] Vladimir Kolmogorov and Yuri Boykov, "An Experimental Comparison of Min-Cut/Max-

Flow Algorithms for Energy Minimization in Vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1124-1137, 2004.

[61] V Goldberg, "A New Max-Flow Algorithm," 1985.

[62] S Wernicke, "Efficient detection of network motifs," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 3, no. 4, pp. 347-359, 2006.

[63] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," 2005.

[64] (2011, May) SocialFlow Blog. [Online]. http://blog.socialflow.com/post/5246404319/breaking-bin-laden-visualizing-the-power-of-a-single

[65] Project Cascade. [Online]. http://nytlabs.com/projects/cascade.html

[66] Truthy. [Online]. http://truthy.indiana.edu/

[67] Jacob Ratkiewicz et al., "Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams," 2010.

[68] Asterisq. [Online]. http://apps.asterisq.com/mentionmap/

[69] Karsten Schmidt and Sascha Pohflepp. Social Collider. [Online]. http://socialcollider.net/

[70] Christian Swinehart. Arbor.js. [Online]. http://arborjs.org/

[71] Megan Mcardle. (2011, May) The Atlantic. [Online]. http://www.theatlantic.com/national/archive/2011/05/anatomy-of-a-fake-quotation/238257/

[72] Martina Morris, Ann Kurth, Deven, Moody, James Hamilton, and Steve Wakefield, "Concurrent Partnerships and HIV Prevalence Disparities by Race: Linking Science and Public Health Practice," *American Journal of Public Health*, vol. 99, no. 6, pp. 1023-1031, 2009.

[73] Ю Ли Куан, *Сингапурская история: из "третьего мира" - в "первый"*.: МГИМО-Университет, 2010.

[74] Ицхак Адизес, *Управление жизненным циклом корпорации*.: Питер, 2008.

[75] Google Visualization API Reference. [Online]. http://code.google.com/intl/ru-

RU/apis/chart/interactive/docs/reference.html#DataTable

[76] Юрий Лифшиц. Структура сложных сетей. Алгоритмы для интернета. [Online]. http://yury.name/internet/04iah.pdf

[77] Google fusion tables. [Online]. http://www.google.com/fusiontables/

[78] Ryan Dahl. NodeJS. [Online]. http://nodejs.org/

[79] Wynn Netherland. (2010, Nov.) The Changelog. [Online]. http://thechangelog.com/post/1489735759/node-canvas-render-and-stream-html5-canvas-using-node-js

[80] Neo4j: The Graph Database. [Online]. http://neo4j.org/

[81] Unrcam.com. [Online]. http://uncram.com

[82] Habrahabr.ru. [Online]. http://habrahabr.ru/

[83] Artjom Kurapov, "Agile web-crawler: design and implementation," Tallinn, 2007.

[84] Remco Van der Hofstad, "Random Graphs and Complex Networks," 2010.

[85] Corey Kosak, Joe Marks, and Stuart Shieber, "Automating the Layout of Network Diagrams with Specified Visual Organization," 1993.

[86] Mozilla people. [Online]. http://people.mozilla.com/~prouget/demos/worker_and_simulatedannealing/index.xhtml

[87] Twitter API. [Online]. http://dev.twitter.com/pages/rate-limiting

[88] NetworkX. [Online]. http://networkx.lanl.gov/examples/drawing/giant_component.html

[89] SocNetV. [Online]. http://socnetv.sourceforge.net/screenshots.html

[90] SocNetV. [Online]. http://socnetv.sourceforge.net/docs/manual.html

[91] Valdis Kerbs. Orgnet.com. [Online]. http://www.orgnet.com/tnet.html

[92] Uncloaking a Slumlord Conspiracy with Social Network Analysis. [Online]. http://www.orgnet.com/slumlords.html

[93] Twitter Network Visualizations: "Stanford". [Online]. http://www.flickr.com/photos/marc_smith/4311427445/in/set-72157622437066929/

[94] Boost C++ Libraries. [Online]. http://www.boost.org/doc/libs/1_46_1/libs/graph/doc/maximum_matching.html

[95] Jerry Neumann. (2010, Nov.) Reaction Wheel. [Online]. http://reactionwheel.blogspot.com/2010/11/venture-coinvestment-map.html

[96] James Moody, "Race, School Integration, and Friendship Segregation in America," *AJS*, vol. 107, no. 3, pp. 679-716, November 2001.

[97] Batagelj V. (2005) NICTA Networks Workshop. [Online]. http://vlado.fmf.uni-lj.si/pub/networks/Doc/Seminar/NICTA.htm

[98] Rose Hoberman and Roni Rosenfeld. (2002) Using WordNet to Supplement Corpus Statistics. [Online]. http://www.cs.cmu.edu/~roseh/Slides/sphinxLunch02-wordnet.pdf

[99] Gephi Blog. [Online]. http://gephi.org/2011/the-egyptian-revolution-on-twitter/

[100] iPhone / iPad. Новости и советы. [Online]. http://www.iphones.ru/iNotes/97971

[101] Graph visualization on the web with Gephi and Seadragon. [Online]. http://gephi.org/2010/graph-visualization-on-the-web-with-gephi-and-seadragon/

[102] David Auber. VACN Challenge Dublin Large dataset. [Online]. http://tulip.labri.fr/TulipDrupal/?q=node/1241

[103] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins, "Microscopic Evolution of Social Networks," 2008.

[104] Gueorgi Kossinets and Duncan Watts, "Empirical Analysis of an Evolving Social Network," *Science*, vol. 311, pp. 88-90, January 2006.

[105] Richard Dawkins, *The Selfish Gene*.: Oxford University Press, 1976.

[106] Truthy. [Online]. http://truthy.indiana.edu/memedetail?id=5&resmin=45&theme_id=1#page=networkGraph

[107] Mentionmapp. [Online]. http://apps.asterisq.com/mentionmap/#user-navalny

[108] Social Collider. [Online]. http://socialcollider.net/

[109] "The structure of the nervous system of the nematode Caenorhabditis elegans".

[110] Anders Sandberg and Nick Bostrom, "Whole Brain Emulation. A roadmap," 2008.

[111] Luis Izquierdo, Segismundo Izquierdo, José Galán, and José Santos. Schelling's model of spatial segregation. [Online]. http://luis.izqui.org/models/schelling/index.html

[112] Arbor. [Online]. http://arborjs.org/

[113] Яндекс Пробки - это твиттер на колёсах. [Online]. http://www.iphones.ru/iNotes/93516