Алгоритмы на графах. Продолжение

www.shumkoff.ru

Немного теории:

Степень или валентность вершины графа — количество рёбер графа G, инцидентных вершине x. При подсчёте степени ребро-петля учитывается дважды

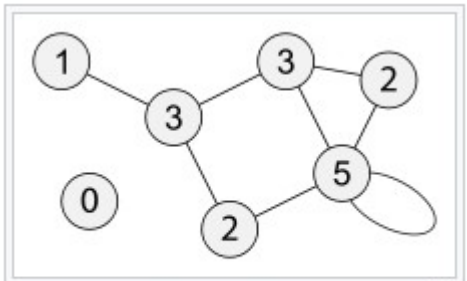
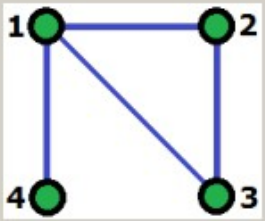


Рис. 1. Граф, на вершинах которого отмечены степени.

Граф



матрица сопряжённости

	1	2	3	4
1		1	1	1
2	1		1	
3	1	1		
4	1			

как правило разреженная

диагональная матрица степеней

	1	2	3	4
1	2			
2		2		
3			2	
4				1

матрица Лапласа

	1	2	3	4
1	2	−1	−1	−1
2	−1	2	−1	
3	−1	−1	2	
4	−1			1

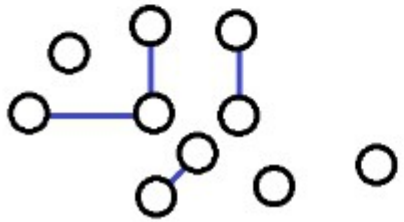
Исследование социальных сетей

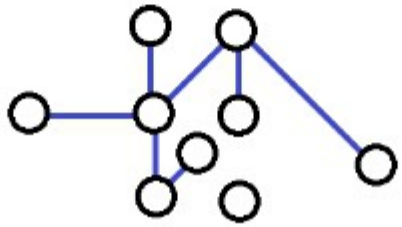
Социальная сеть – динамический граф (пример: мобильная сеть)

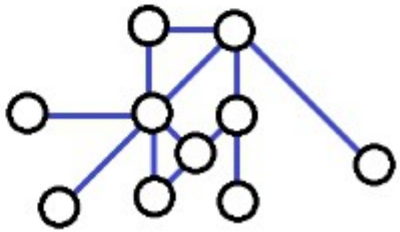
Вершины – пользователи (и группы)

Рёбра – дружба (членство) / связи, отношения

Кластеры – сообщества





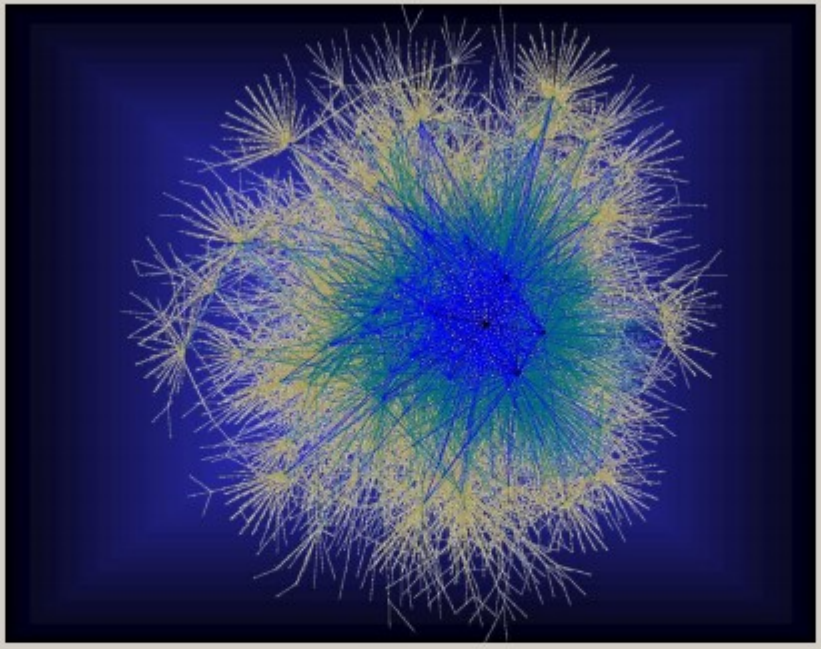


Примеры социальных сетей:
-«классические интернет – социальные сети» (VK, Одноклассники, Facebook, LinkedIn);
-Мобильные сети;
-Научные сообщества (связь по публикациям);
-Почтовые связи (связь по отправке писем);
-Интернет – магазины (связь по купленным //одинаковым?// товарам);
-И сам Интернет (связь по URL и др.)

Какие здесь графы?

Какие задачи здесь актуальны?

Картинки с графами

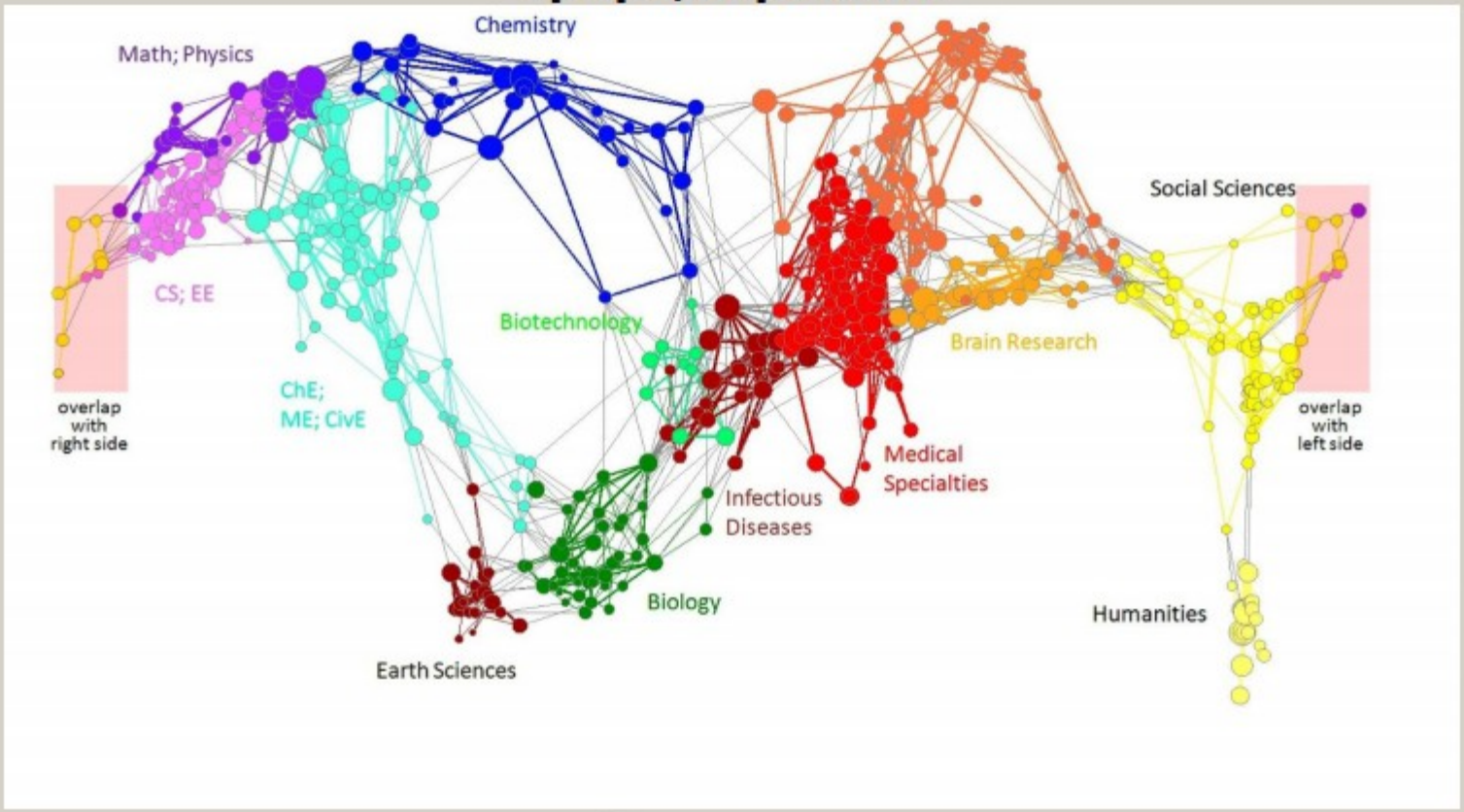


graph of the BGP (Gateway Protocol) web graph, consisting of major Internet routers (6400 вершин, 13000 рёбер)

Ross Richardson, Fan Chung Graham

Примеры графов

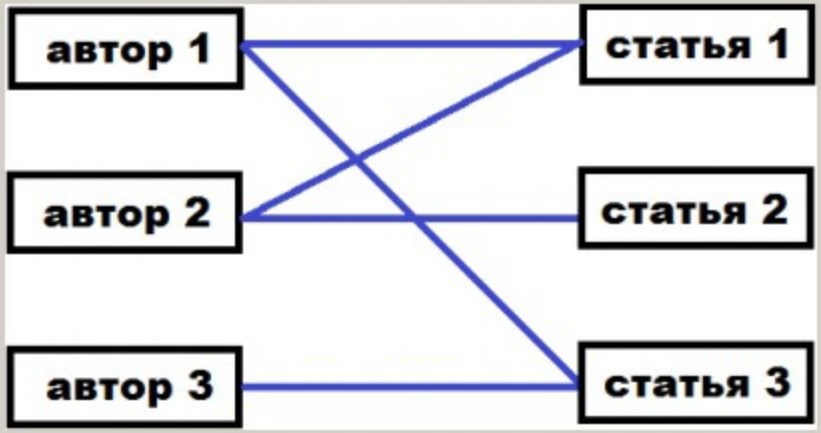
Граф цитирований



Börner и др.

Основные понятия теории графов

Двудольные графы



Научные сообщества

Граф цитирования (ориентированный)

Граф соавторства (неориентированный/двудольный)

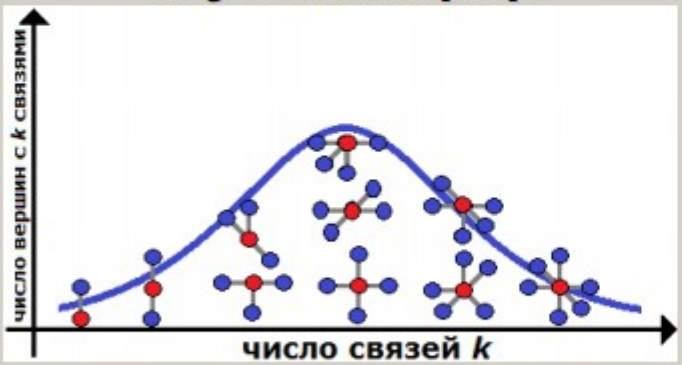
Граф сходства статей (с весами)

1. Распределение вершин

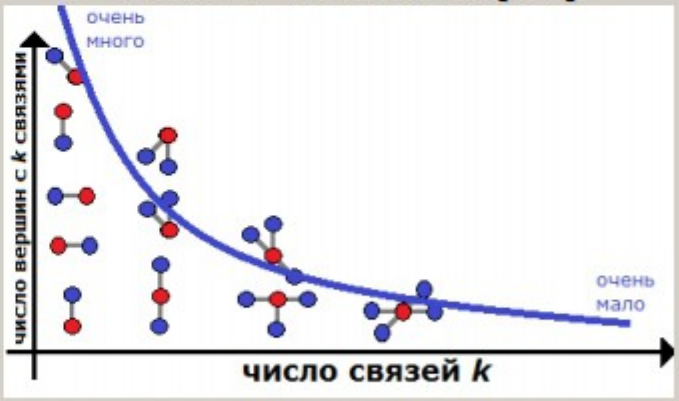
Безмасштабные (scale-free) сети – сети, в которых степени вершин распределены по степенному закону:

доля вершин с k связями ~ $k^{-\gamma}$, обычно $2 < \gamma < 3$.

Случайный граф

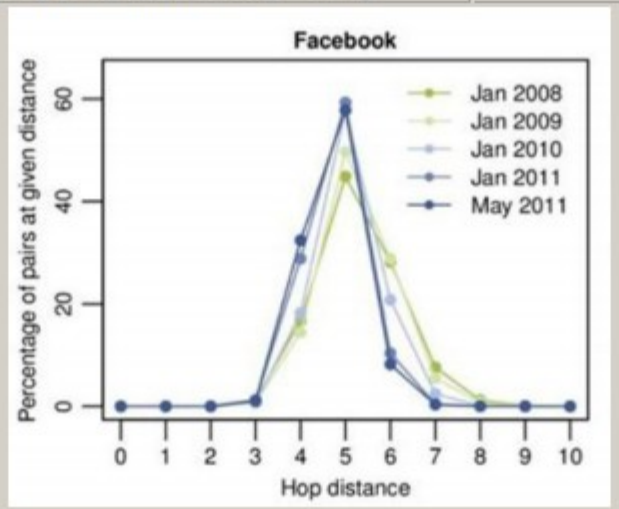


Безмасштабный граф



M.E.J. Newman Power laws, Pareto distributions and Zipf's law // Contemporary Physics, pages 323–351, 2005

Граф	Среднее расстояние между вершинами
Граф почтовых рассылок (D. Watts, 2001, 48000 вершин)	**6**
Граф сообщений в MSN Messenger (J. Lescovec и др. 2007, 240 млн. вершин)	**6.6**
Граф Фейсбука (L. Backstrom и др. 2012, 720 млн. вершин)	**4.74**

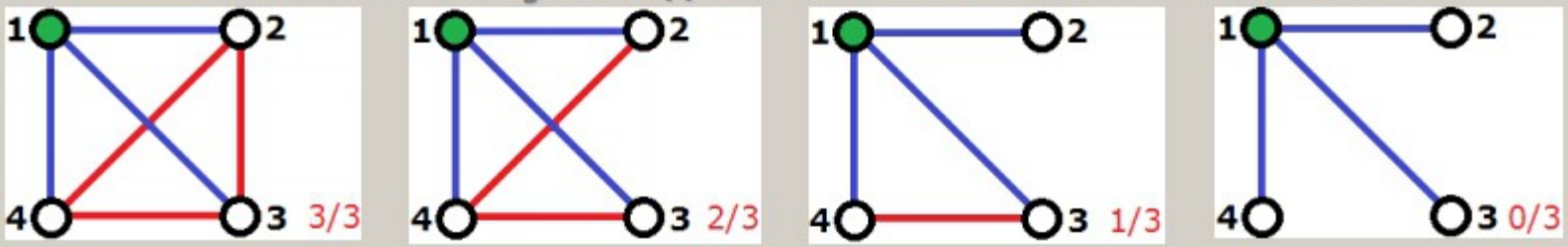


3. Коэффициент кластеризации (clustering coefficient)

2. Локальный

для вершины = насколько её соседи близки к образованию клики

число связей у соседей / число возможных связей



Кликой неориентированного графа называется подмножество его вершин, любые две из которых соединены ребром. Клики являются одной из основных концепций теории графов и используются во многих других математических задачах и построениях с графами. Клики изучаются также в информатике — задача определения, существует ли клика данного размера в графе

Признак графа – функция от признаков вершин (рёбер, …)

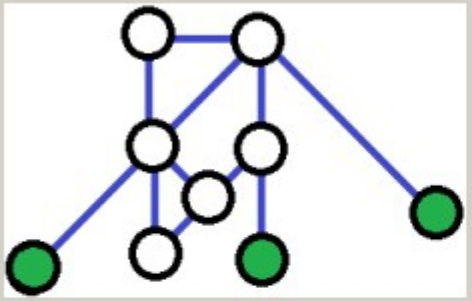
Любая функция!

- **сумма**
- **среднее**
- **максимум**
- **минимум**
- **медиана**
- **сумма квадратов**

и т.п.

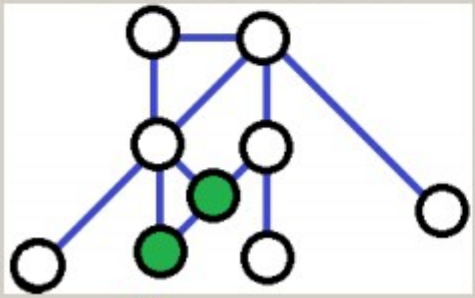
Сходство вершин

1. Формальная (по характеристикам)



По информации о членах соцсети: в одной группе института, одни интересы, участвовали в одном мероприятии

2. По близости



Два близких друга, близнецы

Как определить эти похожести на практике?

Варианты:
-Выделить признаки на графе
-Оценка расстояний

Какие вершины считать важными?:
-По отдельным признака (много соседей)
-- важна вершина соединена с важной

Задачи с социальными сетями

Анализ поведения пользователей:

1) **Выявление аккаунтов – дубликатов;**
2) **Выявление нарушителей;**

Прогнозирование:

1) **Прогнозирование поведения пользователей (какими услугами будет пользоваться, в какие группы будет вступать, с кем «подружится»);**
2) **Прогнозирование и предотвращение ухода пользователя (из группы, сети, …)**
3) **Предсказание траффика**

Рекомендации:

1) **Предсказание эффективности рекламы**
2) **Формирование таргетированных предложений (рекламы, заполнения профиля,**
3) **вступлению в группы)**

Интересная терминология

Степенная центральность (Degree centrality) – число соседей

Центральность по близости (Closeness centrality) – $\sum_{u \neq v} \frac{1}{d(u,v)}$

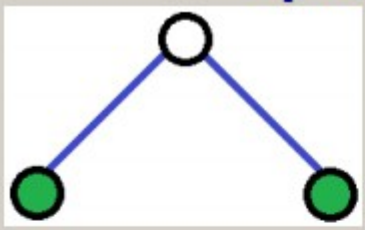
Центральность по путям (Betweenness centrality) – число кратчайших путей, проходящих через эту вершину

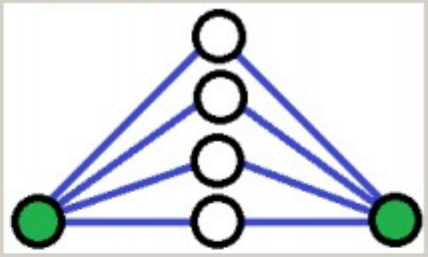
Собственная центральность (Eigenvector centrality) –
центральность вершины зависит от центральности соседей

$$c_i = \sum_j a_{ij} c_j$$

Источник – Дьяконов А.Г. (МГУ)

признак №1 – число соседей





Чем больше общих друзей имеют Иван и Пётр, тем более вероятней, что они подружатся.

$|\Gamma(x) \cap \Gamma(y)|$ **– хорошая мера сходства вершин, где** $\Gamma(x)$ **– множество соседей вершины** x

признак №2

$|\Gamma(x)| \cdot |\Gamma(y)|$ **– коэффициент предпочтительности**

Чем более общительны, тем скорее подружатся

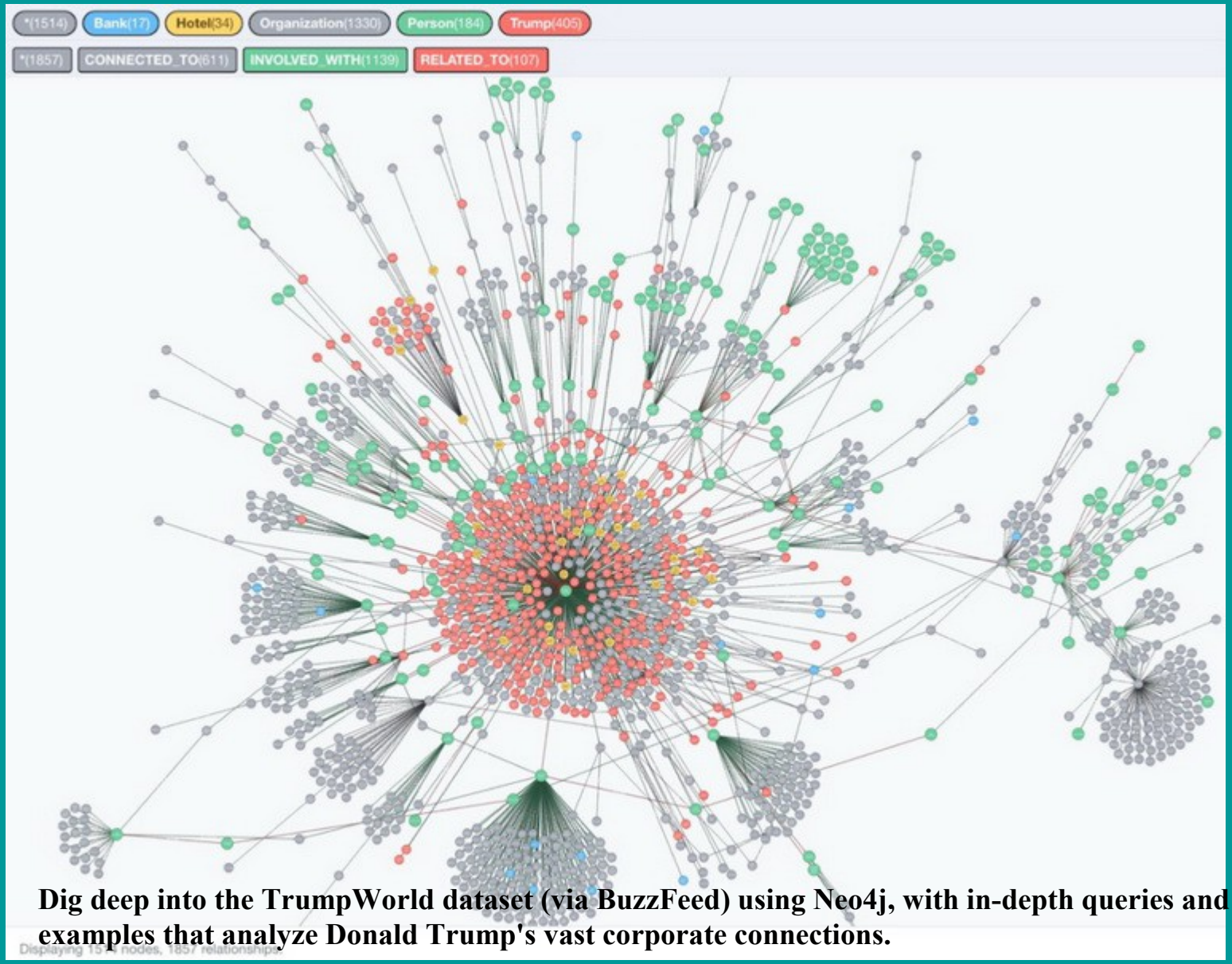
Программный инструментарий для работы с графами

Графовые СУБД:
-СУБД «Neo4j»
-«OrientDB»
-- СУБД «Titan»

Модель neo4j предусматривает два типа сущностей: вершины, которые могут обладать набором свойств, и ребра, связывающие эти вершины, при этом ребра также могут обладать набором свойств и могут относиться к разным типам.

Orient DB представляет собой комбинированную графово-документную базу, позволяющую хранить в узлах не просто набор свойств, а целые документы с динамической схемой.

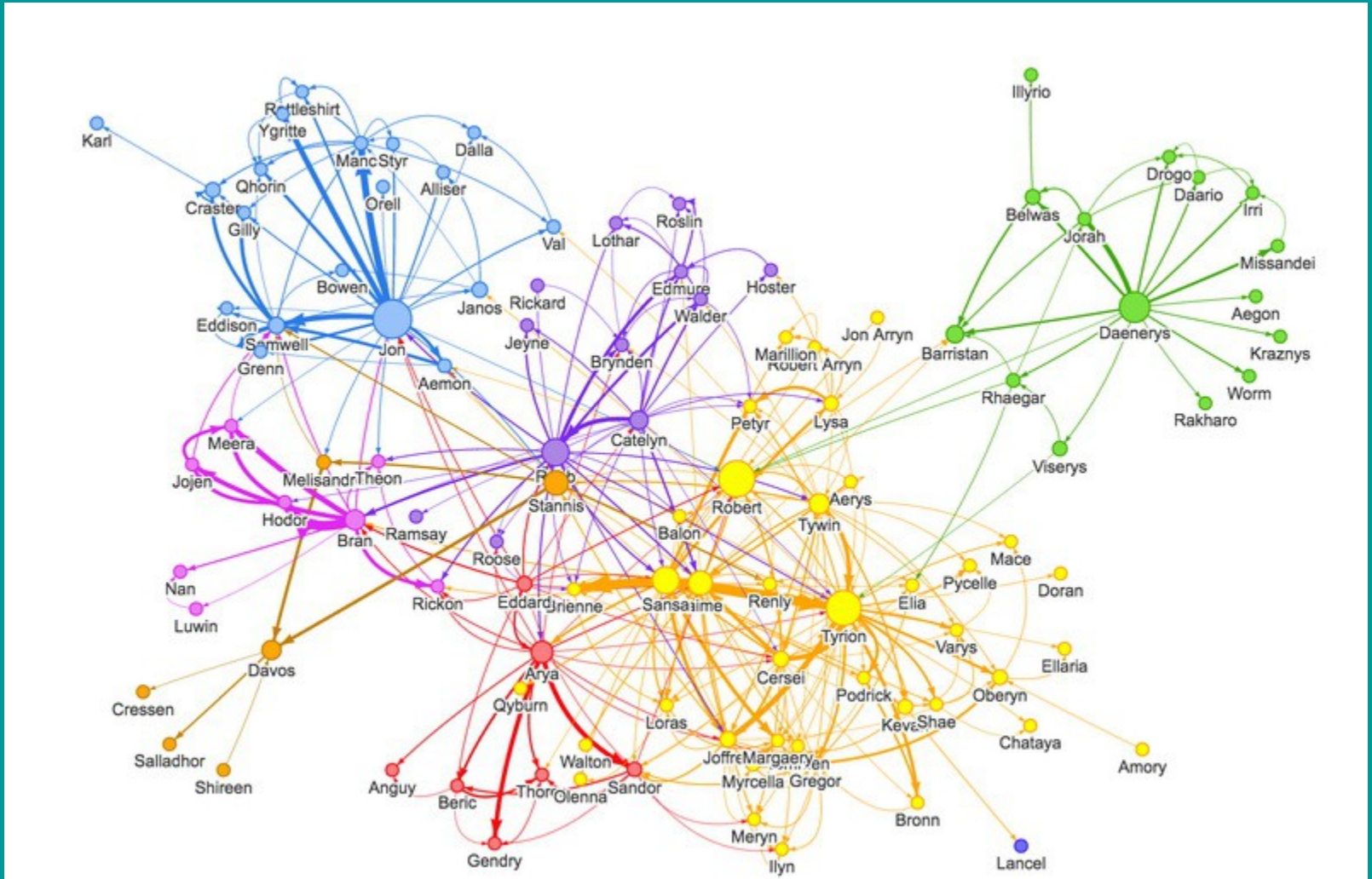
Пример работы СУБД Neo4j



Dig deep into the TrumpWorld dataset (via BuzzFeed) using Neo4j, with in-depth queries and examples that analyze Donald Trump's vast corporate connections.

https://www.pinterest.co.uk

Пример работы СУБД Neo4j



Социальные связи + кол-во сообщений

Источник – medium.com