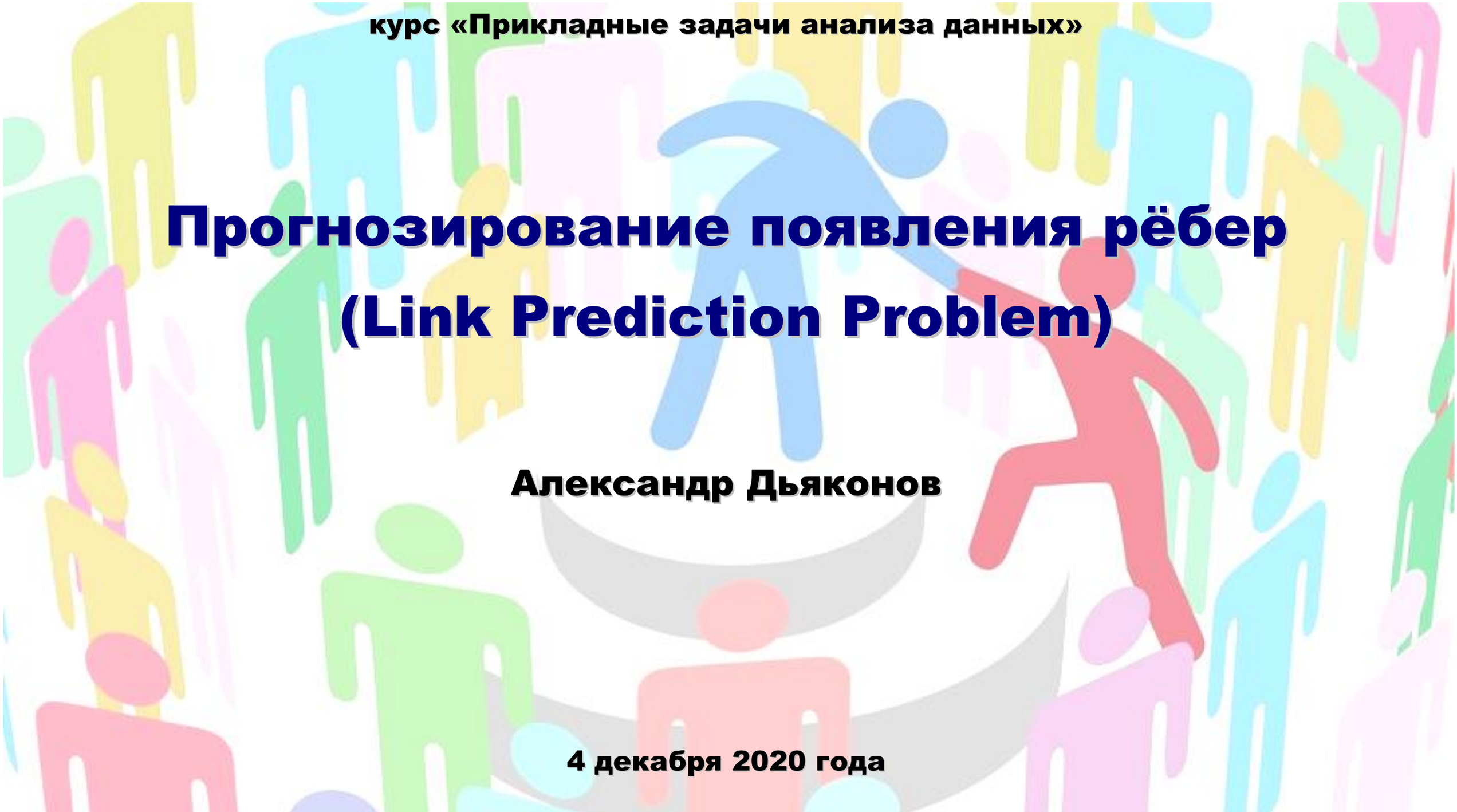
курс «Прикладные задачи анализа данных»



Прогнозирование появления рёбер (Link Prediction Problem)

Александр Дьяконов

4 декабря 2020 года

План

Признаковые пространства, построенные по графам

Сходство вершин

Важность вершин

PageRank и его модификации

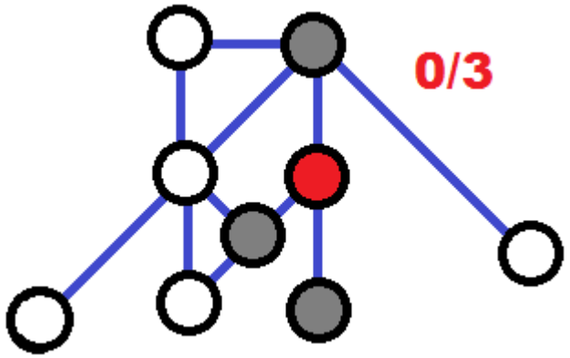
case: соревнование

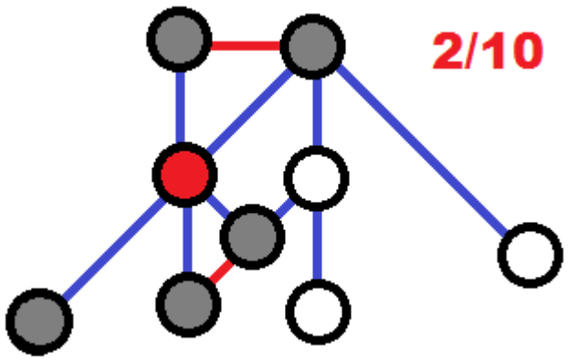
Часто графы просто погружают в признаковое пространство…

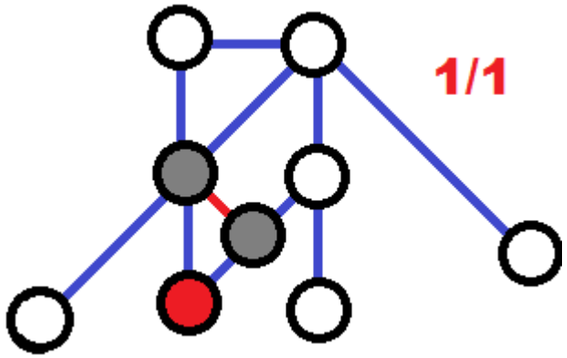
и граф превращается в вектор

Пример признака (уже был)

Коэффициент полноты (clustering coefficient)







что такое 10?

характеризует полноту эго-графа одной вершины

(~ окрестность первого порядка)

Как интерпретировать?

В чём недостаток?

Как исправить?

Недостатки

лучше использовать в сочетании с другими признаками (например, число соседей)

Это типично для признаков на графе!

Как придумать признак для всего графа (а не отдельной вершины)?

Как придумать признаки для всего графа

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

Любая функция! Любая статистика!

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

и т.п.

Сходство вершин

Часто надо измерить сходство двух вершин/рёбер/подграфов

Какие бывают похожести?

Что значит, что вершины похожи?

Важность вершин

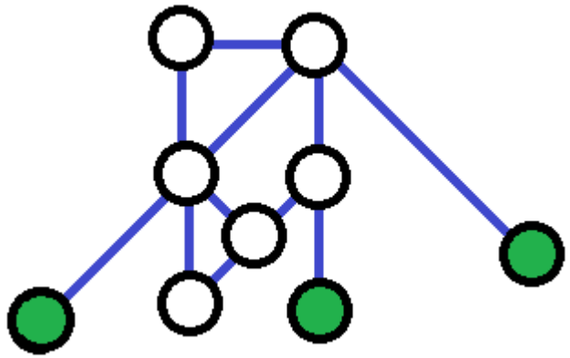
Часто надо измерить особенность вершины/ребра/подграфа

Например, для поиска непохожих вершин, влиятельных блогеров

Какие вершины считать «важными»?

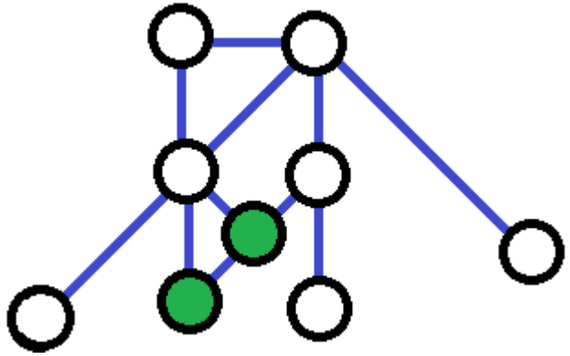
Сходство вершин

1. Формальная
(по характеристикам)



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

2. По близости

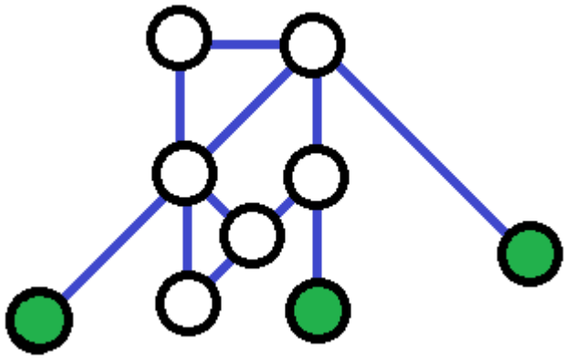


Два близких друга, близнецы

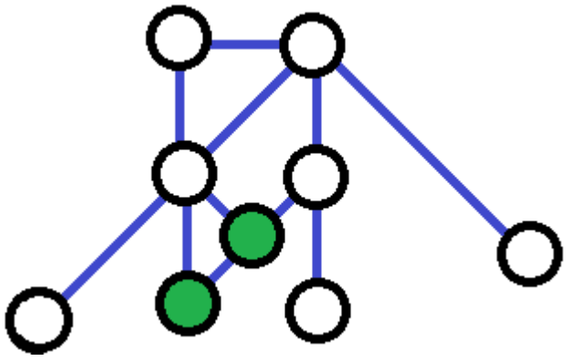
Как определить эти похожести на практике?

Сходство вершин

1. Формальная
(по характеристикам)



2. По близости



Как измерить?

Погружение в признаковое пространство

Вычисление сходства в нём

Оценка расстояния на графе

Важность

Какие вершины считать важными?

- По отдельным признакам (например, много соседей)
- По рекурсии (важная вершина соединена с важными)

Пример важности – центральность вершины (сейчас рассмотрим)

Кстати, а что такое граф? С точки зрения реализации

Очень полезно

Любой объект имеет много представлений
(подпространство, многогранник и т.п.)

1. **С точки зрения определения**

2. **С точки зрения реализации**

Разреженная матрица

Объекты (пользователи) – строки/столбцы

Аппарат линейной алгебры

3. **С точки зрения сути**

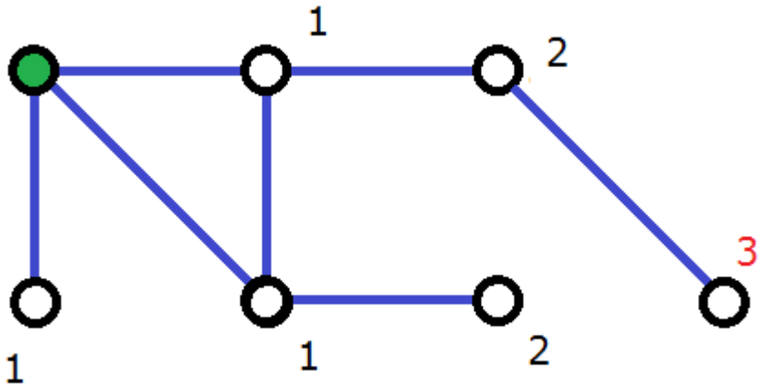
Это формализация отношений

Важны окрестности большого порядка, их свойства, связи,
не всё может быть отражено в графе

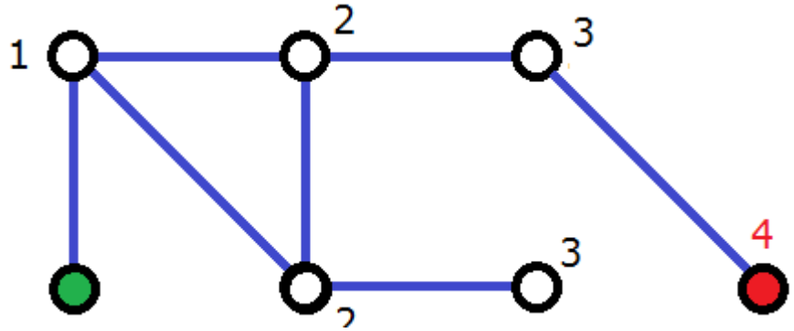
Центральность вершины в графе

Эксцентриситет – **вершины** v

$$\varepsilon(v) = \max_{u \in V} d(u,v)$$

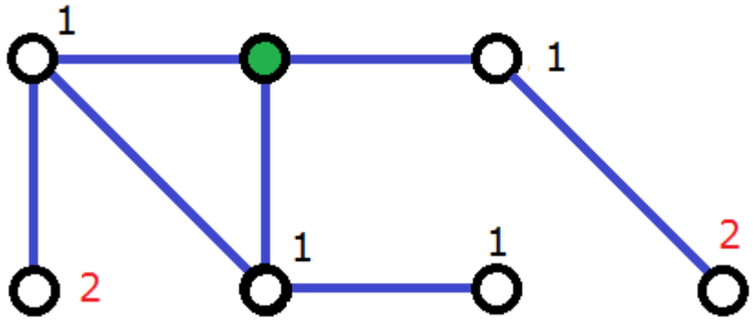


Диаметр – **максимальный эксцентриситет**



Центральность вершины в графе

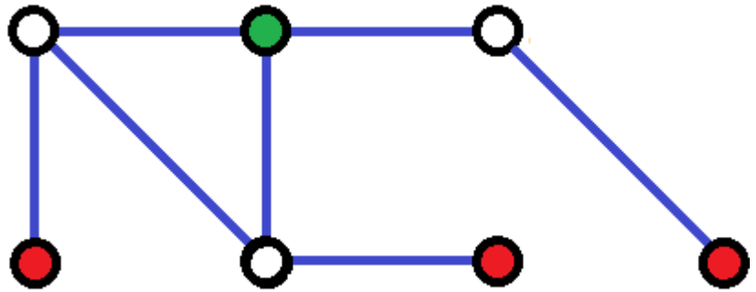
Радиус – минимальный эксцентриситет



Вершина графа центральная,
если её эксцентриситет равен радиусу графа.

Центр – множество центральных точек

Периферия – множество точек с максимальным эксцентриситетом



Интересная терминология

Степенная центральность (Degree centrality) – число соседей

$$C_{\text{degree}}(i) = d_i$$

$$d_{\text{out}} = A\tilde{1}$$
$$d_{\text{in}} = A^{\text{T}}\tilde{1}$$
$$a_{ij} \sim (i \to j)$$

ij-й элемент ~ дуга из i в j

Быстрое вычисление: $O(1)$

Normalizing Degree Centrality

$$C^{\text{norm}}_{\text{degree}}(i) = \frac{d_i}{n_v} \vee \frac{d_i}{\max\limits_j d_j} \vee \frac{d_i}{2n_e}$$

$$\sum d_j$$

Центральность по близости (Closeness centrality) –

$$\sum_{u \neq v} \frac{1}{d(u,v)}$$

нужны все попарные расстояния

алгоритм Дейкстра

$$O(n_v^2 \log n_v + n_v n_e)$$

предполагается связность графа

иногда:

$$\frac{1}{\frac{1}{n_e - 1} \sum_{i \neq j} d(i,j)}$$

Центральность по путям (Betweenness centrality)

– число (доля) кратчайших путей, проходящих через эту вершину

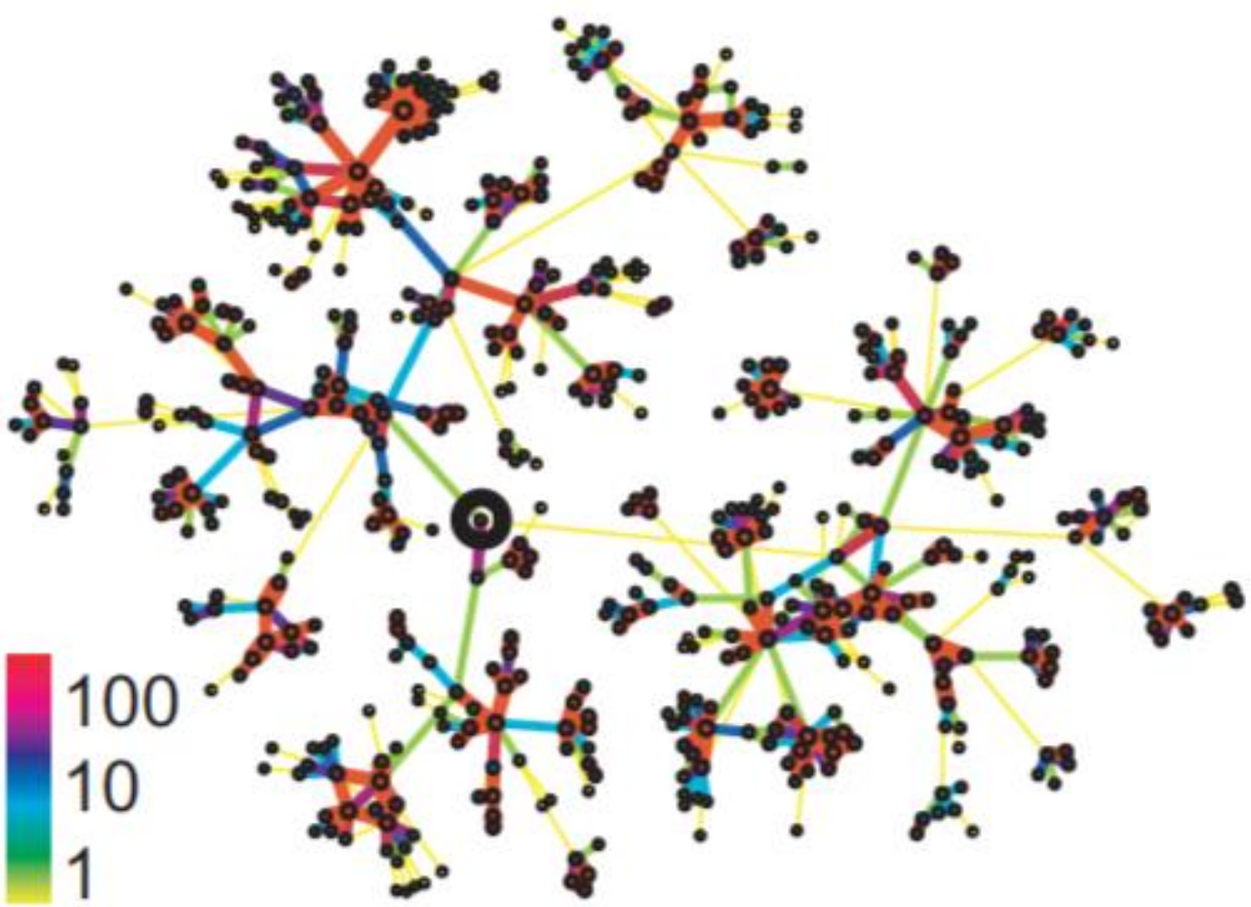
$$C_{\text{betweenness}}(i) = \frac{\#\text{paths}(s \to t, i)}{\#\text{paths}(s \to t)}, \quad \begin{matrix} s \neq t \neq i \\ s \neq t \neq i \end{matrix}$$

Центральность ~ если ходить по графу,
то часто посещаешь эту вершину

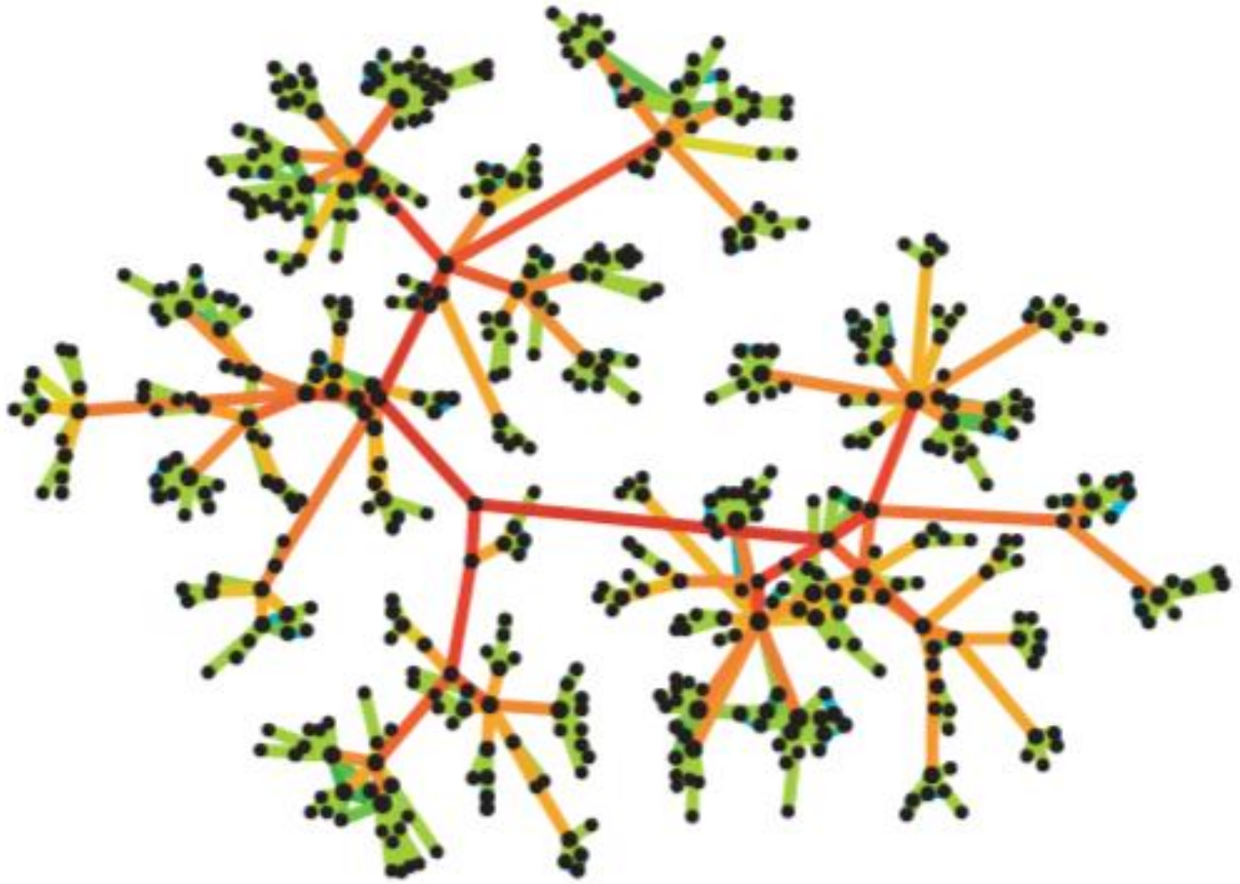
Есть $O(n_v n_e)$ алгоритм (для графа без весов)

U. Brandes «A faster algorithm for betweenness centrality» // Journal of Mathematical Sociology, vol. 25, no. 2, pp. 163-177, 2001

Центральность по путям (Betweenness centrality)



Edge strength



Edge betweenness

http://www2.ece.rochester.edu/~gmateosb/ECE442/Slides/block_4_sampling_modeling_inference_part_a.pdf

Собственная центральность (Eigenvector centrality) –

центральность вершины зависит от центральности соседей

$$c_j \propto \sum_i a_{ij} c_i$$

$$(D^{-1}A)x = x$$

(тут если по-другому строить матрицу смежности)

max с.з. = 1

собственный вектор ~ max с.з.

Метод:

- **вычисление собственных векторов**
- **взятие вектора с максимальным собственным значением**
- **его значения – центральности вершин**

дальнейшая модификация ~ см. PageRank

Katz –

взвешенная сумма путей, приходящих в вершину

Путь длины k берём с коэффициентом β^k, $\beta \in [0,1]$

$$(\beta A + \beta^2 A^2 + \beta^3 A^3 + \ldots)\tilde{1} =$$
$$(\beta A + \beta^2 A^2 + \beta^3 A^3 + \ldots)(I - \beta A)(I - \beta A)^{-1}\tilde{1} =$$
$$(\beta A + \beta^2 A^2 + \beta^3 A^3 + \ldots - \beta^2 A^2 - \beta^3 A^3 - \ldots)(I - \beta A)^{-1}\tilde{1} =$$
$$\beta A (I - \beta A)^{-1}\tilde{1}$$

(тут если по-другому строить матрицу смежности)

На основе этого вычисляется центральность.

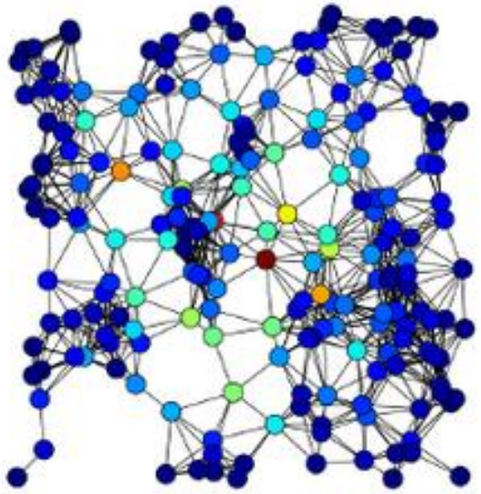
Эксцентриситетная центральность (Eccentricity centrality)

$$e(v) = \frac{1}{\max\limits_u d(u,v)}$$

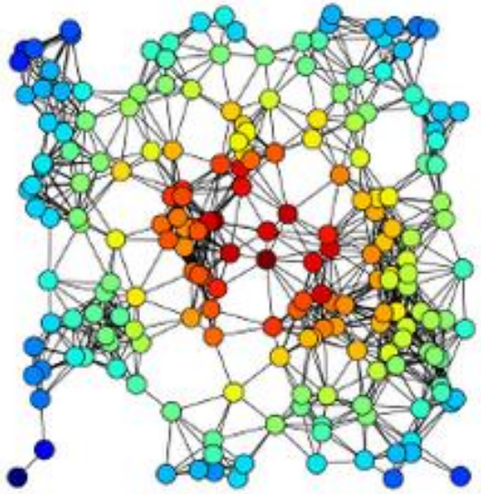
Сложность как в центральности по близости (Closeness centrality)

F.W. Takes and W.A. Kosters, Computing the Eccentricity Distribution of Large Graphs, Algorithms, vol. 6, nr. 1, pp. 100-118, 2013

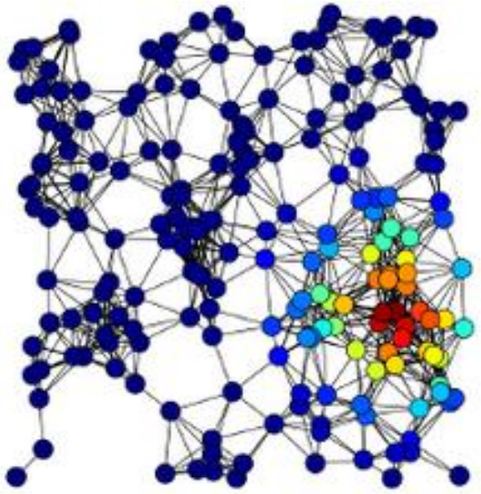
Разные виды центральности



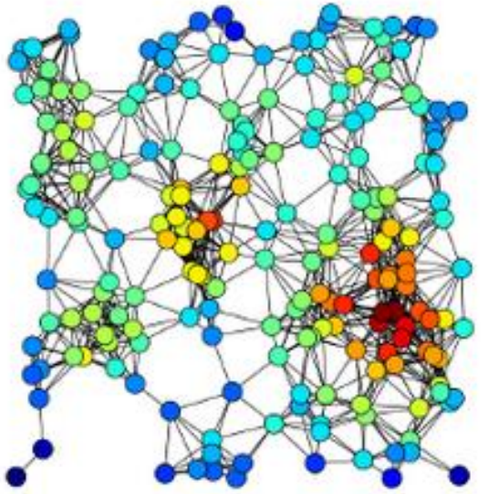
Betweennes centrality



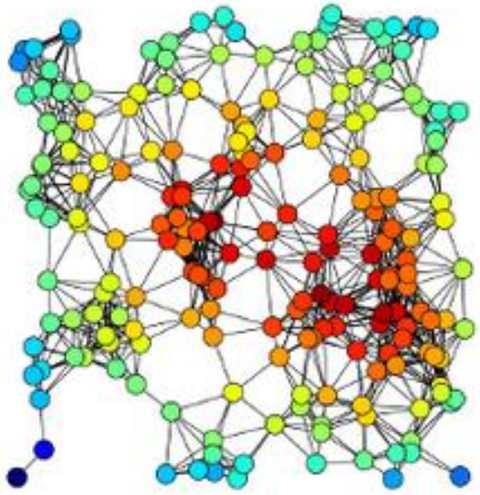
Closeness centrality



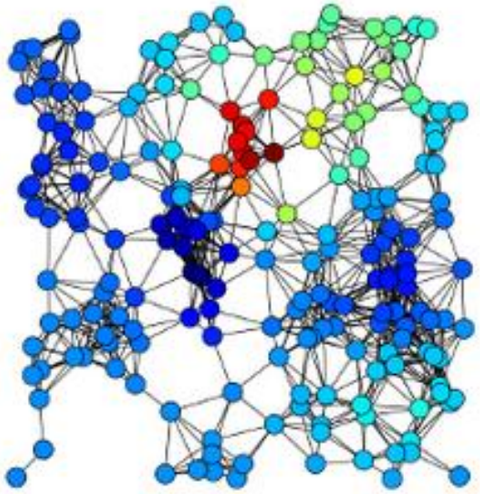
Eigenvector centrality



Degree centrality

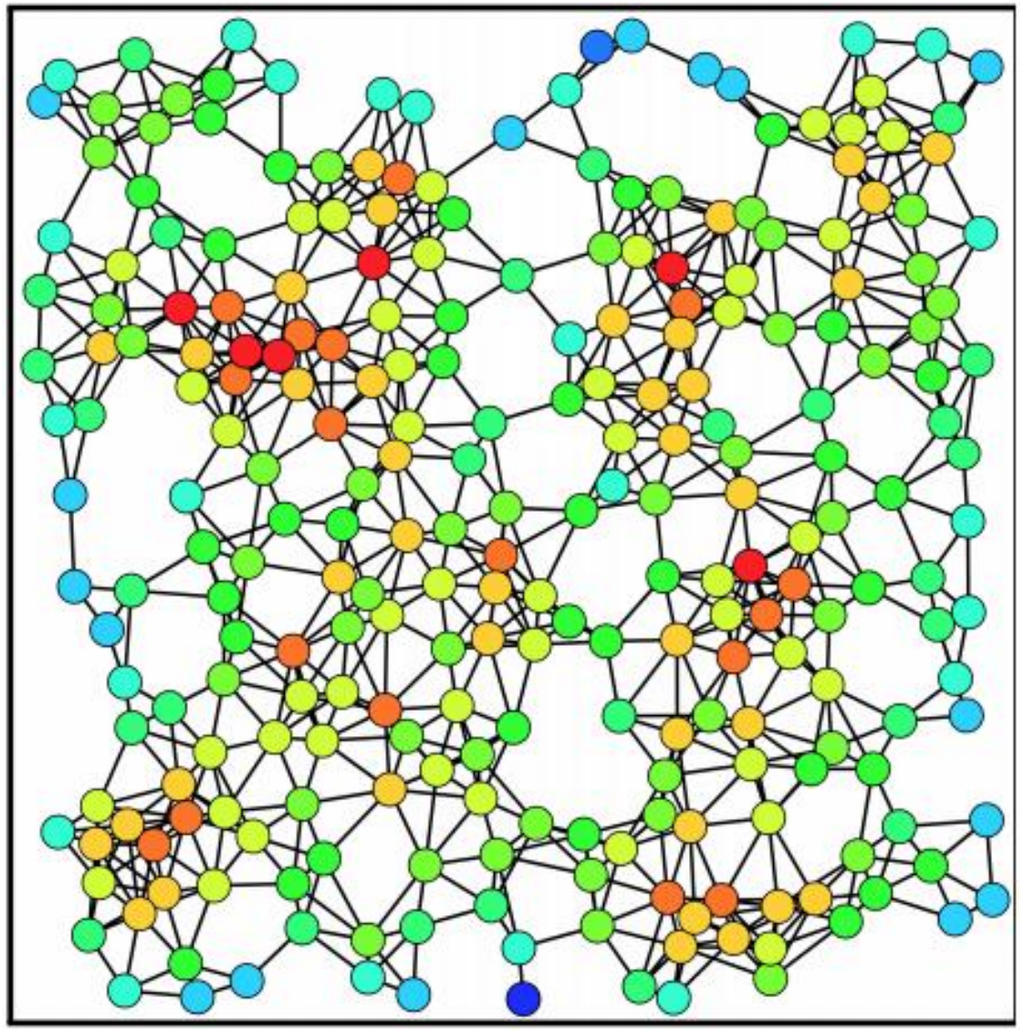


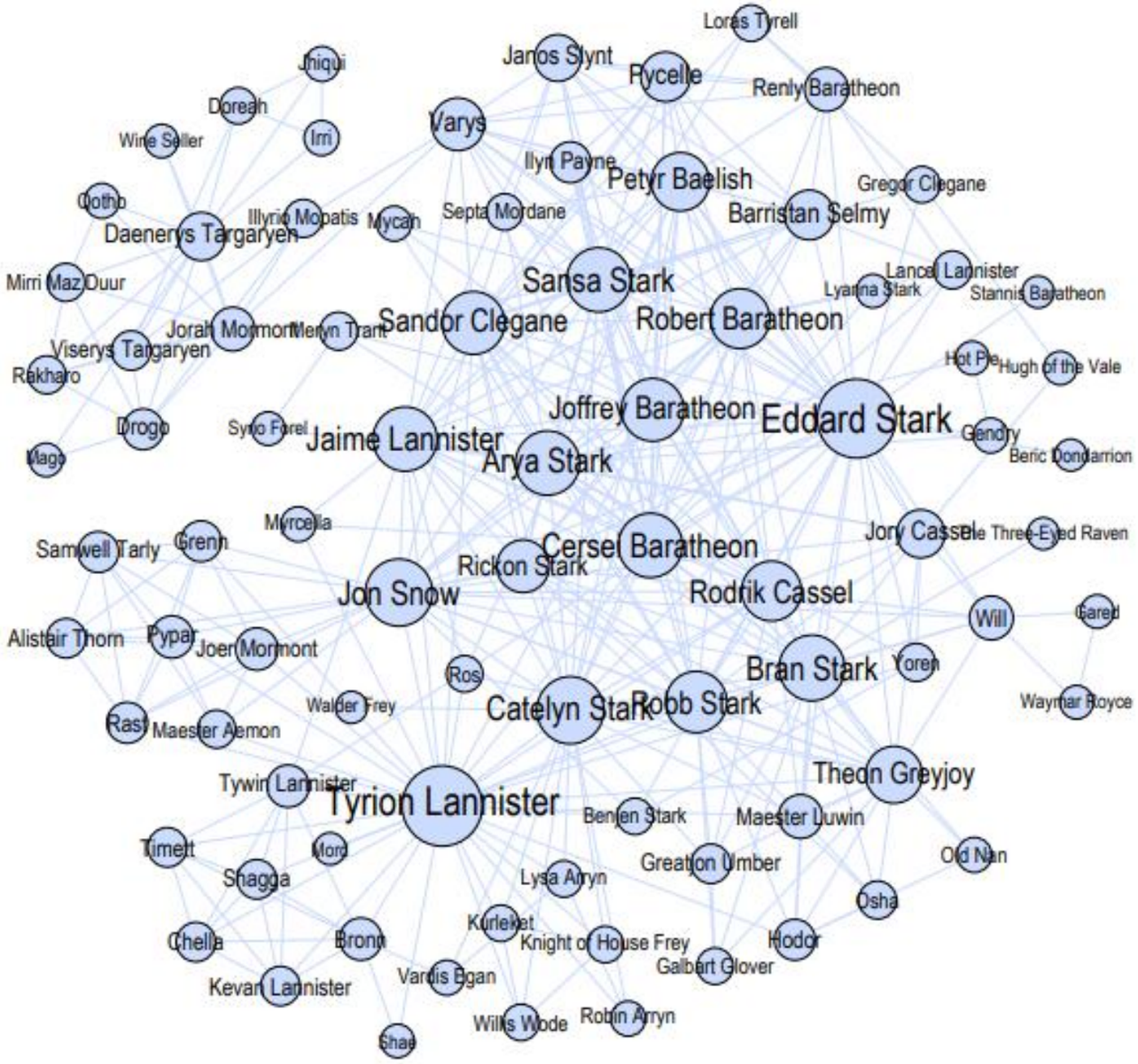
Harmonic centrality



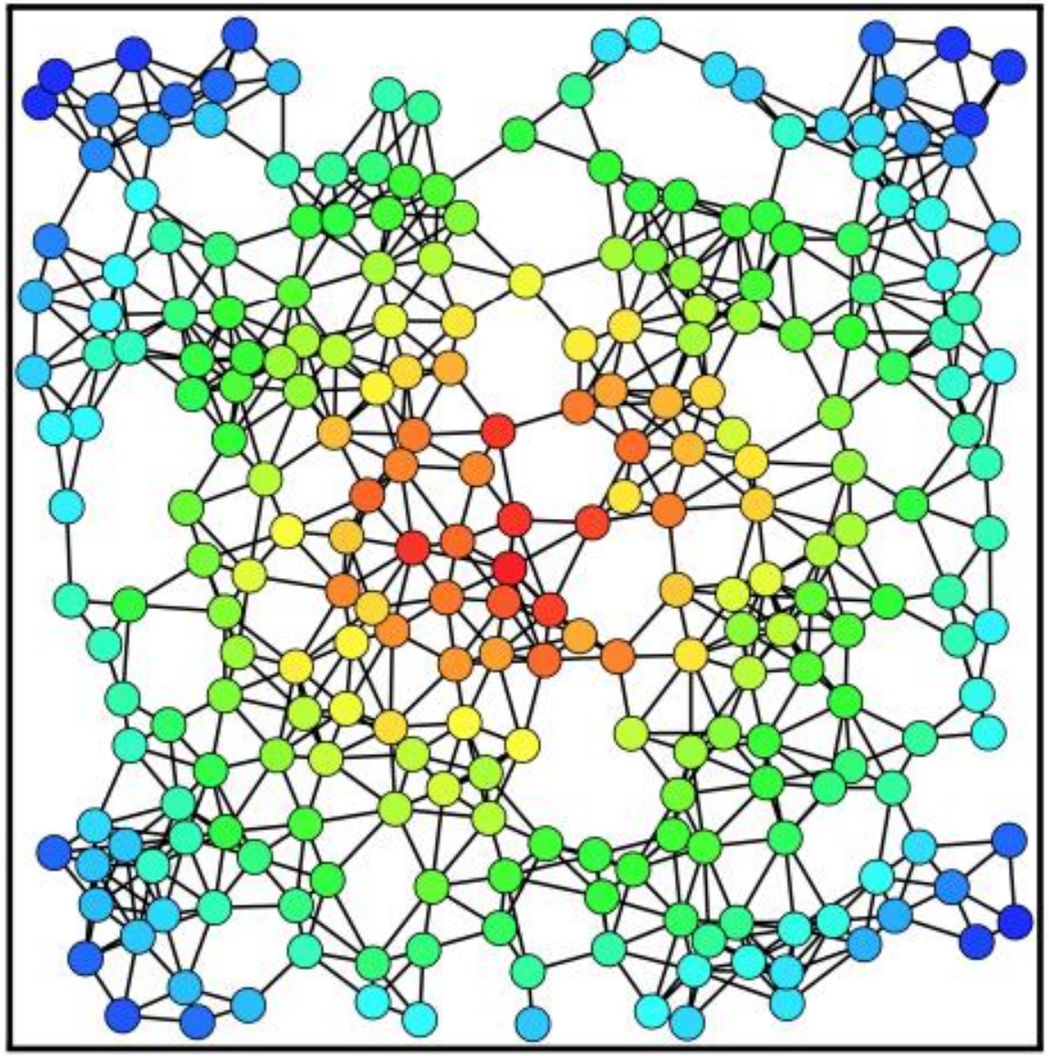
Katz centrality

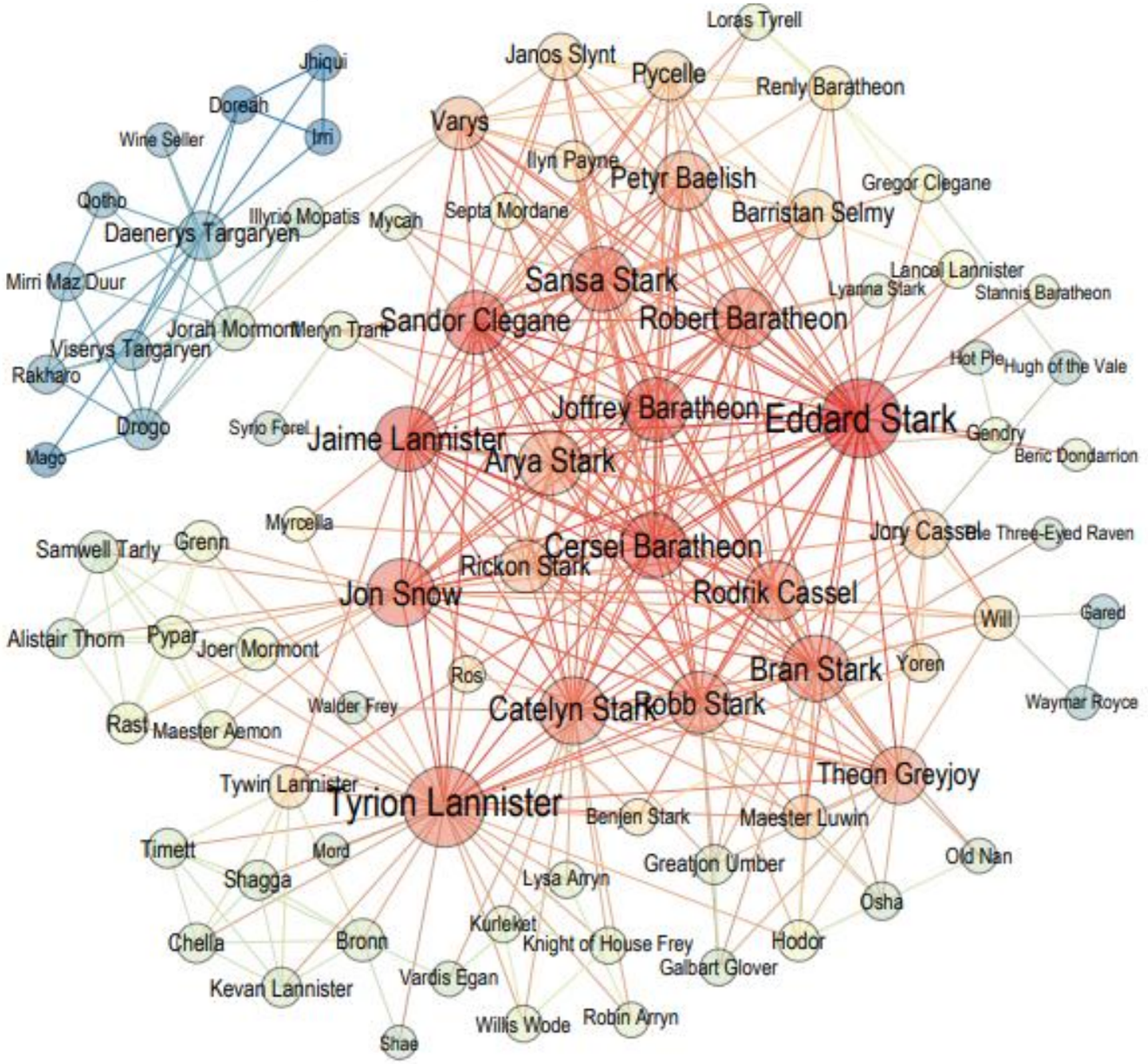
Degree Centrality



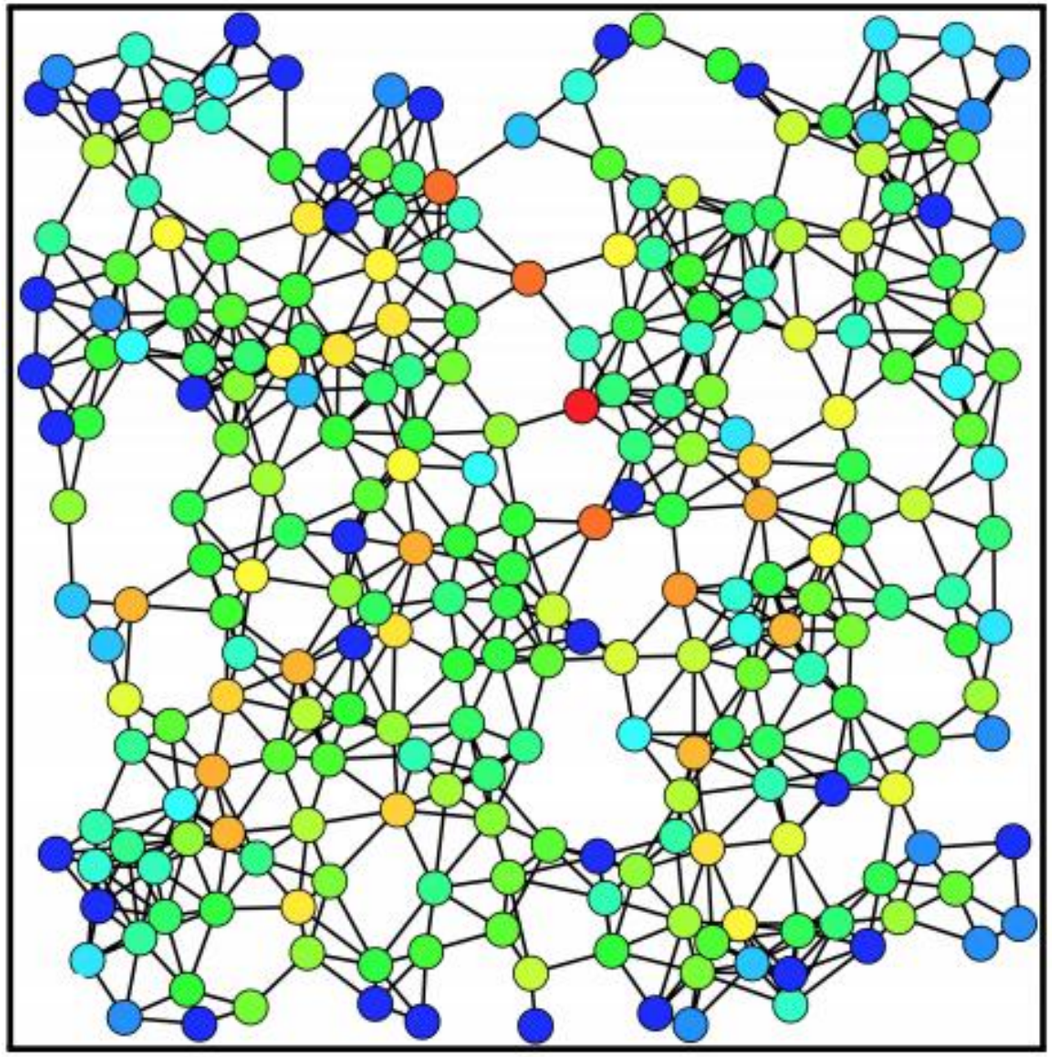


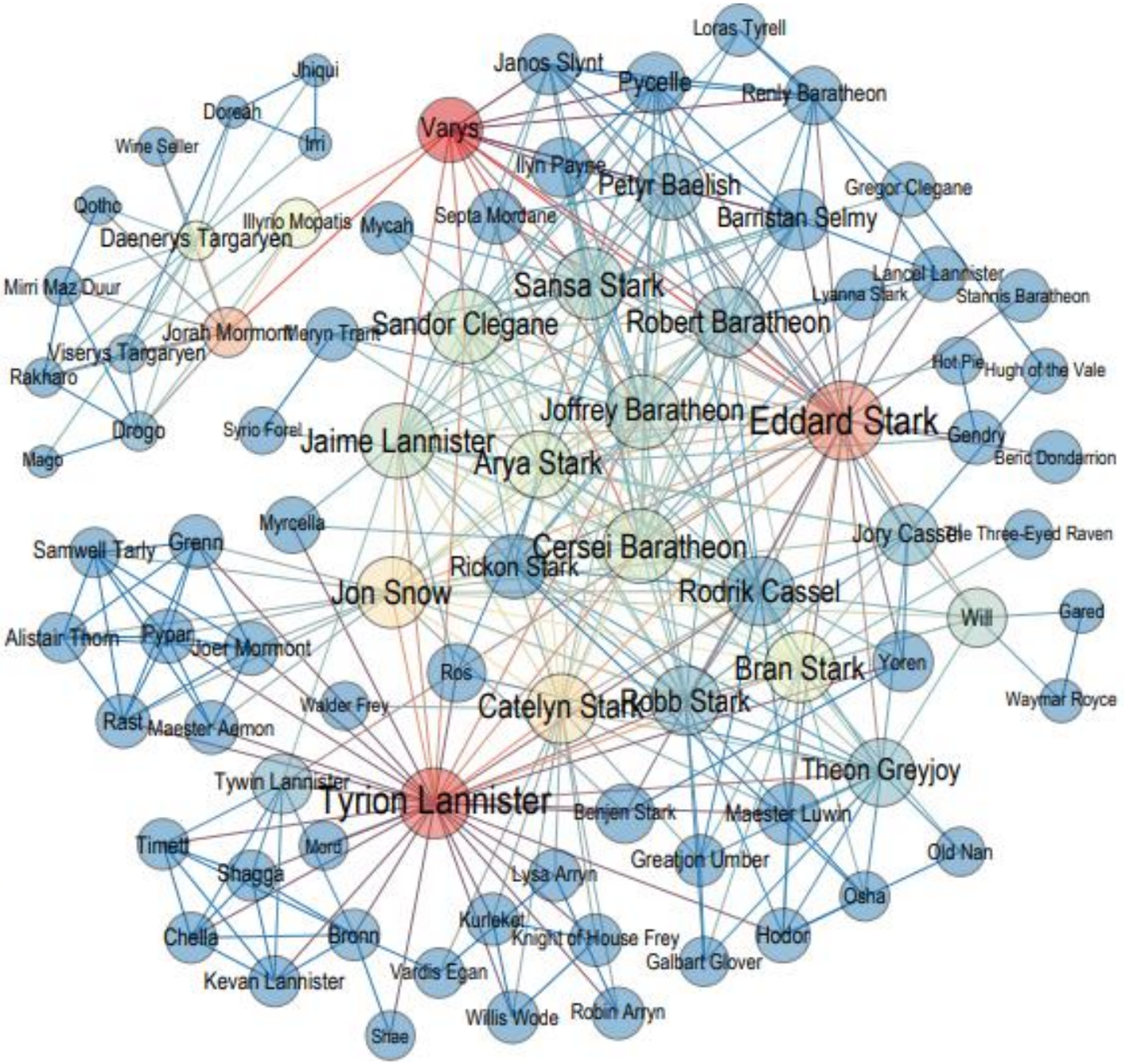
Closeness Centrality



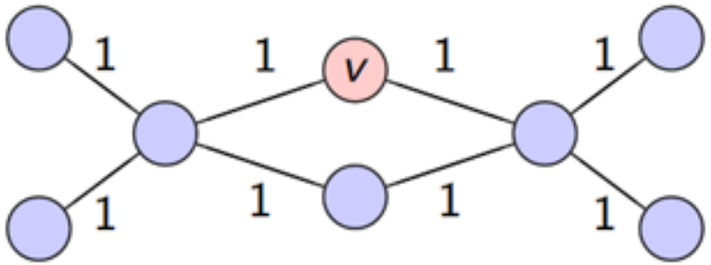


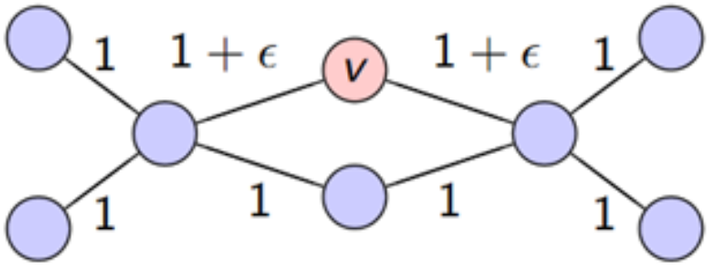
Betweeness centrality



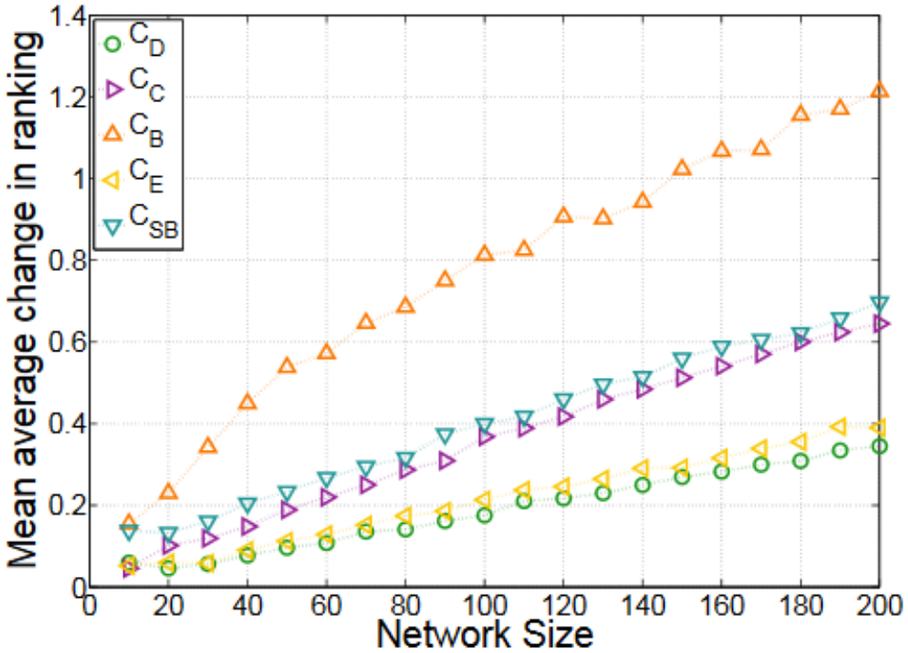


Устойчивость понятий





betweenness centrality неустойчива, но есть устойчивые модификации…



Сравниваются ранки в исходном и чуть подпорченном графе (веса рёбер умножаются)

D = degree
C = closeness
B = betweenness
SB = stable betweenneess

S. Segarra and A. Ribeiro «Stability and continuity of centrality measures in weighted graphs» // IEEE Trans. Signal Process, 2015

Важность группы (Group Centrality)

приведённые понятия легко обобщаются, например

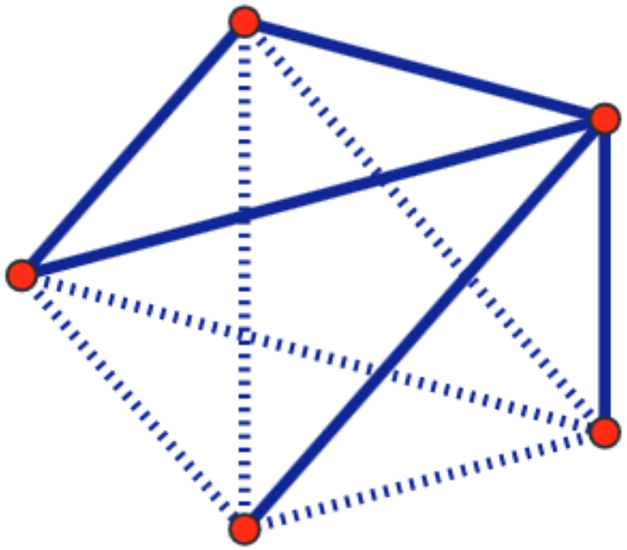
$$C_{\text{degree}}(S) = |\{(i,j) \in E \,|\, i \in S, j \notin S\}|$$

Прогнозирование появления ребра в динамическом графе (Link Prediction Problem)

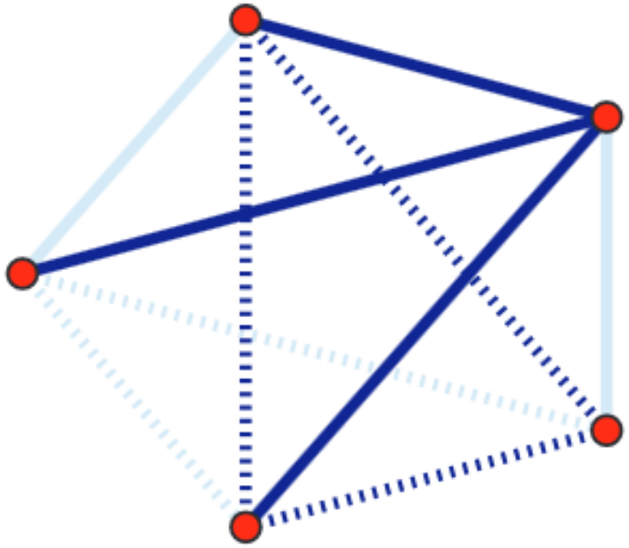
Дан слепок графа соцсети

Какие рёбра появятся в ближайшем будущем?

Чаще: для конкретных пар вершин «вероятность стать ребром»



Original graph



Link prediction

Liben-Nowell et. al. «The link-prediction problem for social networks» // J. of American society for info science and technology. 2007 https://www.cs.cornell.edu/home/kleinber/link-pred.pdf

Прогнозирование появления ребра в динамическом графе (LPP)

Приложения:

социальные сети,

сотовые операторы,

мобильные операторы и т.д.

Как решать?

Для каждой пары (i, j) выпишем потенциально хорошие признаки

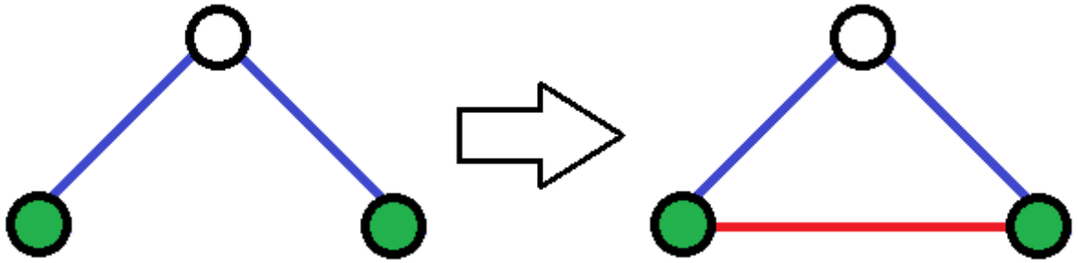
меры схожести вершин

– формирование признакового пространства

признак №0 – расстояние на графе (graph distance)

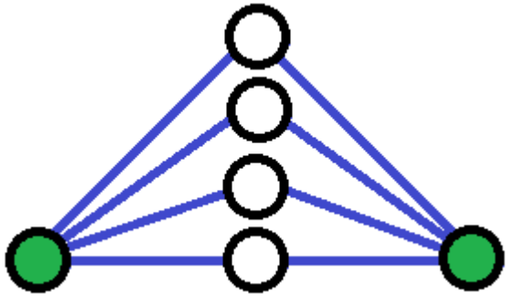
LPP признаки: 1) число соседей (common neighbors)

Принцип
«друг моего друга»



если (x,z) **– ребро,** (z,y) **– ребро,**
то (x,y) **– ребро или станет ребром**

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

В его чём недостатки?

LPP признаки: 2) коэффициент предпочтительности

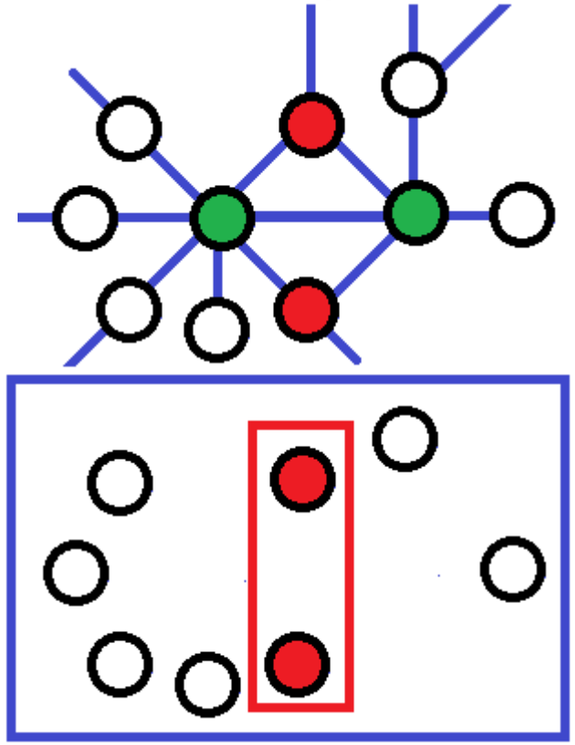
$|\Gamma(x)| \cdot |\Gamma(y)|$ – **коэффициент предпочтительности (preferential attachment)**

Чем более общительны, тем скорее подружатся

LPP признаки: 3) коэффициент Жаккара

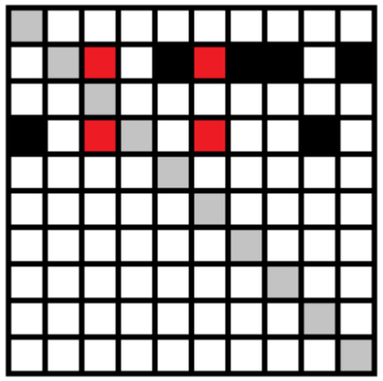
Или наоборот: чем больше процент общих друзей

$$\frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$ **– коэффициент Жаккара (Jaccard's coefficient)**



обычные признаки
для сравнения множеств

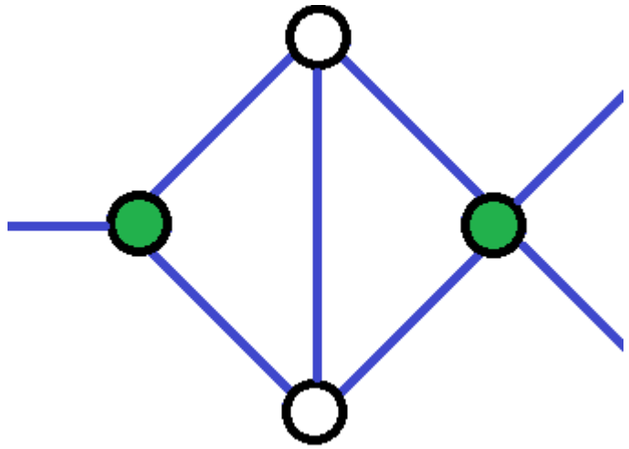
просто сравнение строк матрицы смежности

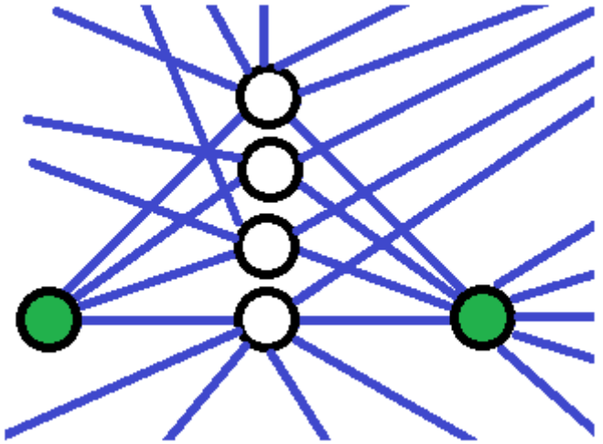


Полезно: разный подход к описанию смысла (множества, строки)

LPP признаки: 4) коэффициент Адамик/Адара

не все друзья одинаковые!

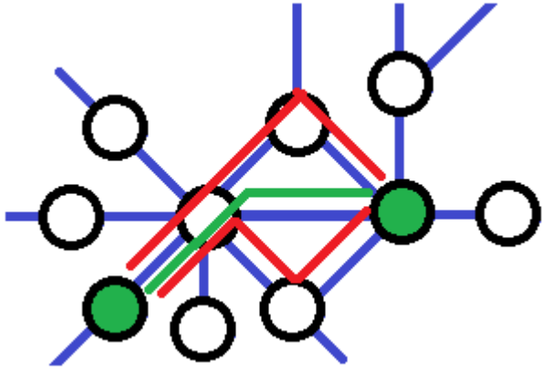




$$\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log|\Gamma(z)|}$$ **– коэффициента Адамик/Адара (Adamic/Adar)**

LPP признаки: 5) Katz

Учитывать целые цепочки друзей-друзей



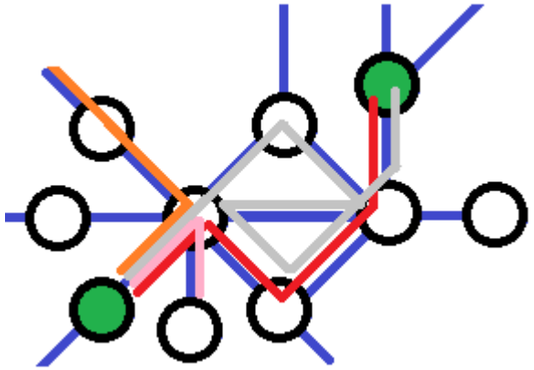
$\sum_{l=1}^{\infty} \beta^l \, \mathrm{path}_l(x,y)$ **– признак Katz**

равен xy-му элементу матрицы

$$(I - \beta M)^{-1} - I,$$

LPP признаки: 6) на основе случайных блужданий

Вершины близки, если из одной легко попасть во вторую



Пример: среднее время достижения вершины

Часто используют не матрицу смежности, а её k-SVD-аналог

+ PageRank

LPP признаки: 7) на основе рекуррентных вычислений

SimRank

Вершины похожи,
если похожи их друзья

$$\mathrm{sim}(x,y) = \frac{\gamma}{|\Gamma(x)|\cdot|\Gamma(y)|}\sum_{a\in\Gamma(x)}\sum_{b\in\Gamma(y)}\mathrm{sim}(a,b)$$

Разные итерации пересчёта можно сделать признаками!

LPP признаки: 8) вероятностные методы

Пусть вершина i порождается с вероятностью $\mathrm{P}(i)$

По ней порождается латентный класс с вероятностью $\mathrm{P}(z|i)$

По нему порождается ребро с вероятностью $\mathrm{P}(j|z)$

Вероятность появления ребра

$$\mathrm{P}(i,j)=\mathrm{P}(i)\cdot\mathrm{P}(z|i)\cdot\mathrm{P}(j|z)$$

– это ответ, вероятности здесь оцениваются EM-алгоритмом, максимизируя логарифм правдоподобия

$$\sum_{\{i,j\}\in E}\log(\mathrm{P}(i,j))$$

Алгоритм PageRank (подробнее про случайные блуждания)

Две эквивалентные интерпретации
«что такое важные страницы в интернете»

I) Случайные блуждания

Если ходить по ссылкам в Интернете,
то важная страница – на которую чаще попадаешь

II) Перетекание рейтинга

«Важные»:
1. На них ссылаются (есть входящие ссылки)
2. На них ссылаются важные страницы

Алгоритм PageRank

Если страница j с важностью w_j имеет $d_{\text{out}}(j)$ выходных ссылок,

каждая ссылка «передаёт» важность

$$\frac{w_j}{d_{\text{out}}(j)}$$

Важность страницы = сумма всех входных ссылок

$$w_j = \sum_{(i,j)\in E} \frac{w_i}{\deg_{\text{out}}(i)}$$

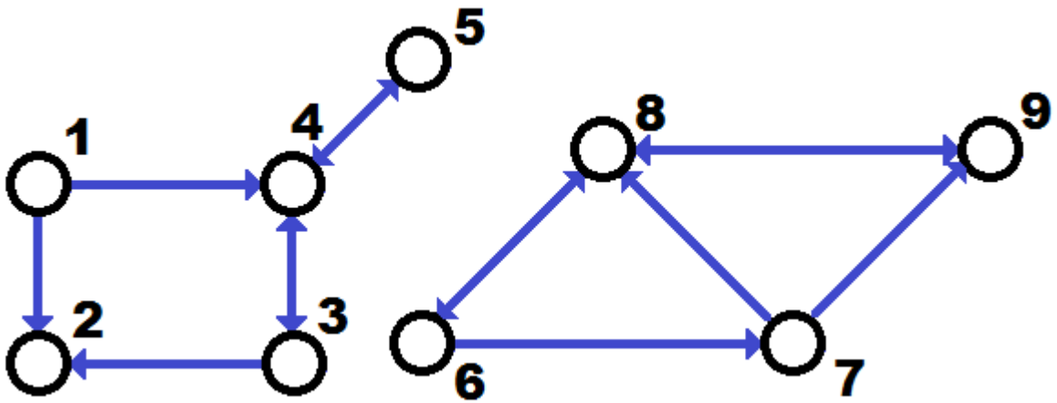
Если пронормировать матрицу смежности

$$N = D_{\text{out}}^{-1} A$$

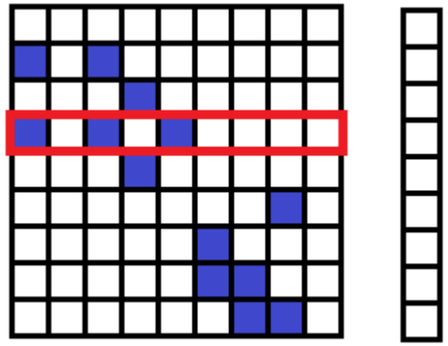
тогда вектор важности рекурсивно записывается как

$$w = N^{\text{T}} w$$

Алгоритм PageRank



тут транспонированная матрица



$$w_4 = \frac{w_1}{2} + \frac{w_3}{2} + \frac{w_5}{1}$$

Внимание на построение матрицы смежности!

Алгоритм PageRank

Решаем задачу на собственные значения

$$N^{\mathrm{T}} w = \lambda w$$

Наибольшее с.з. = 1

Берём его собственный вектор!

Итерационный метод

$$w^{(t)} = N^{\mathrm{T}} w^{(t-1)}$$

это и находит

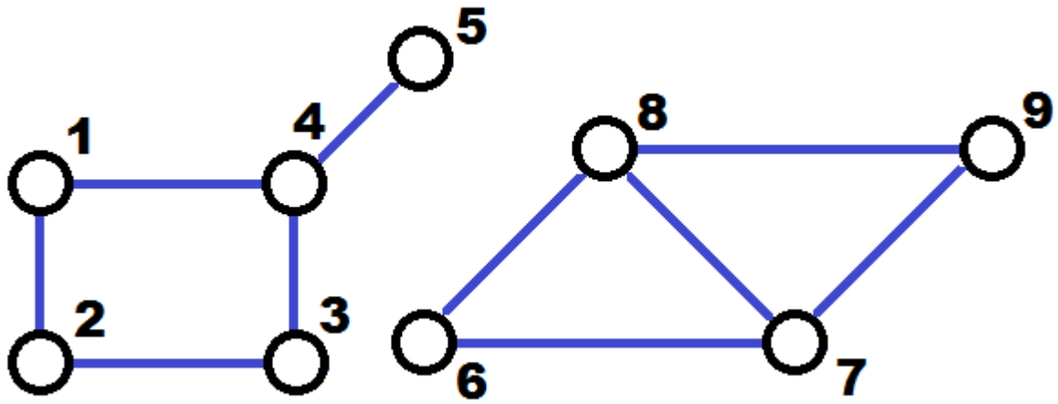
Алгоритм PageRank

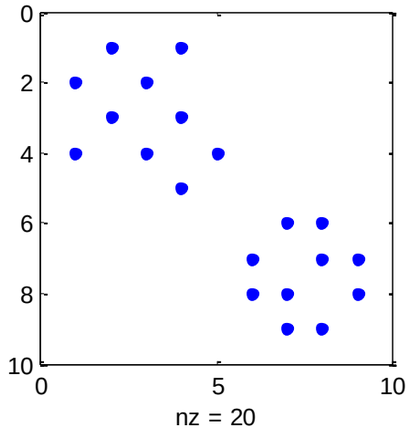
Можно по-разному формализовать

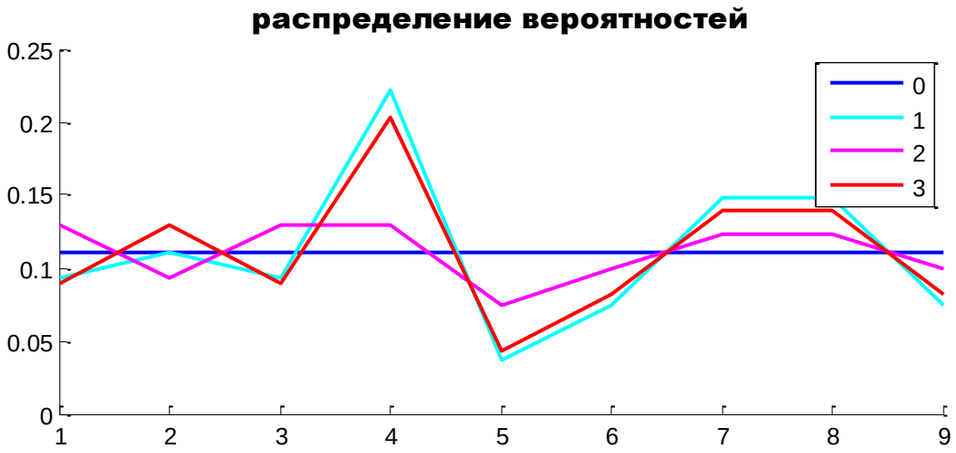
Если матрицу отнормировать, то сумма рангов в сети – константа

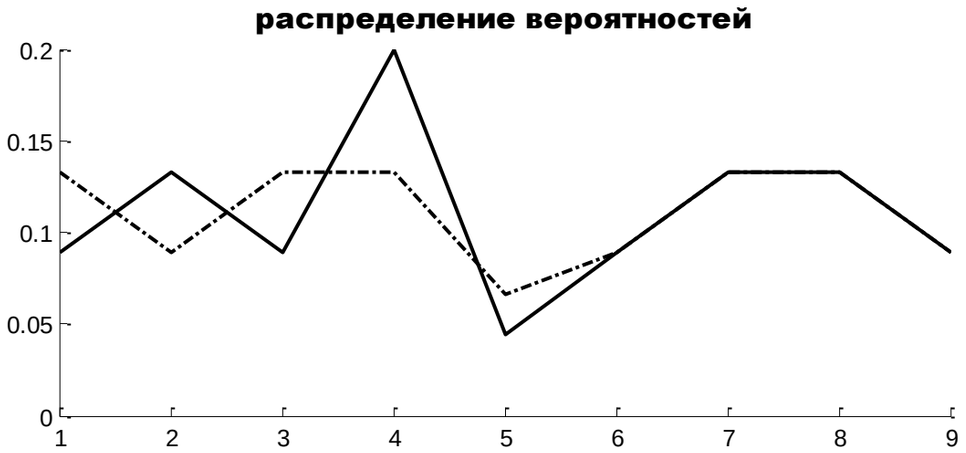
+) рейтинг не появляется, он постоянен в сообществе

Алгоритм PageRank: проблема на практике









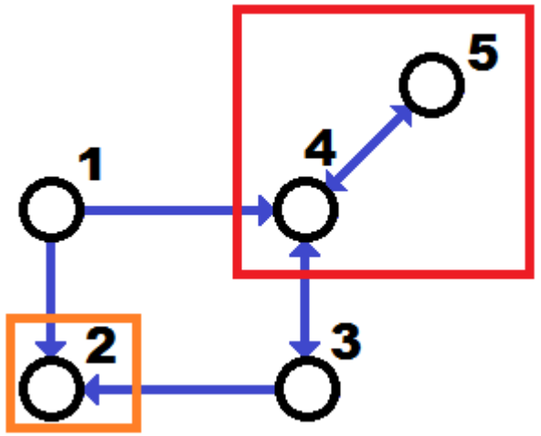
не всегда получается

Почему?

Алгоритм PageRank: два типа проблем

1. Циклы (Spider traps)

2. Мёртвые вершины (Dead ends)



Решение: в итерационном алгоритме с вероятностью 0.1-0.2 прыгать в случайную вершину графа (~5 шагов)

Алгоритм PageRank: решение проблем

Брин, Пейдж:

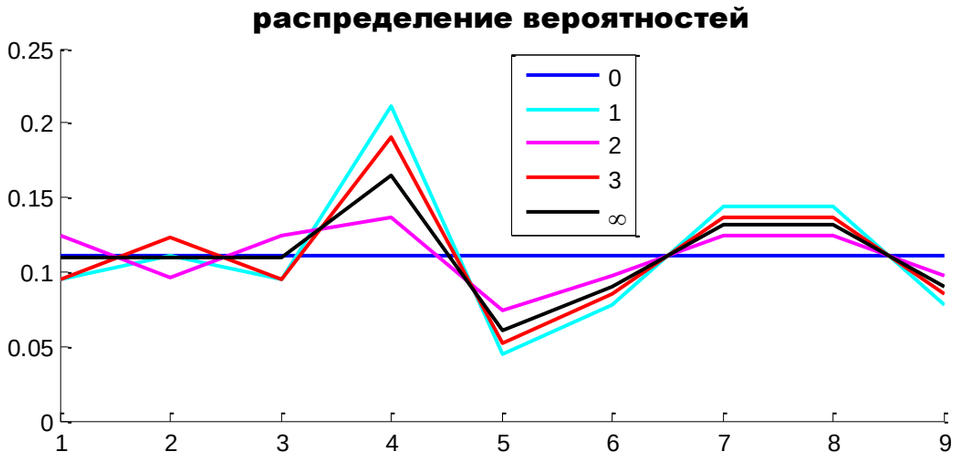
$$w_j = \beta \sum_{(i,j)\in E} \frac{w_i}{\deg_{\text{out}}(i)} + (1-\beta)\frac{1}{n}$$

$$M = \beta \cdot N + \frac{(1-\beta)}{n}\tilde{1}\cdot\tilde{1}^{\text{T}}$$

Обычно 100 итераций

Larry Page and Sergey Brin, The PageRank citation ranking: Bringing order to the web, Technical Report, Stanford Infolabs, 1999. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Алгоритм PageRank: результат



Модификации PageRank: практические аспекты

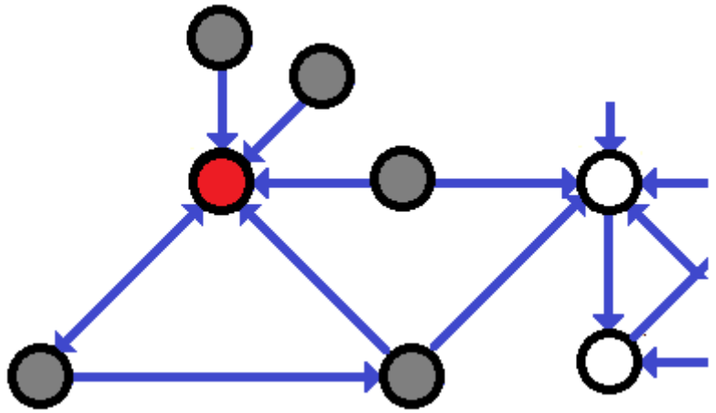
Переход не в произвольную вершину, а

- **в похожую,**
- **из этого топика,**
- **из доверительного множества (анти-спам: *.edu),**
- **в эту вершину (SimRank)**

  и т.п.

Зачем?

Ответ: в случае спама – борьба с фермами спама



– доверенная зона,

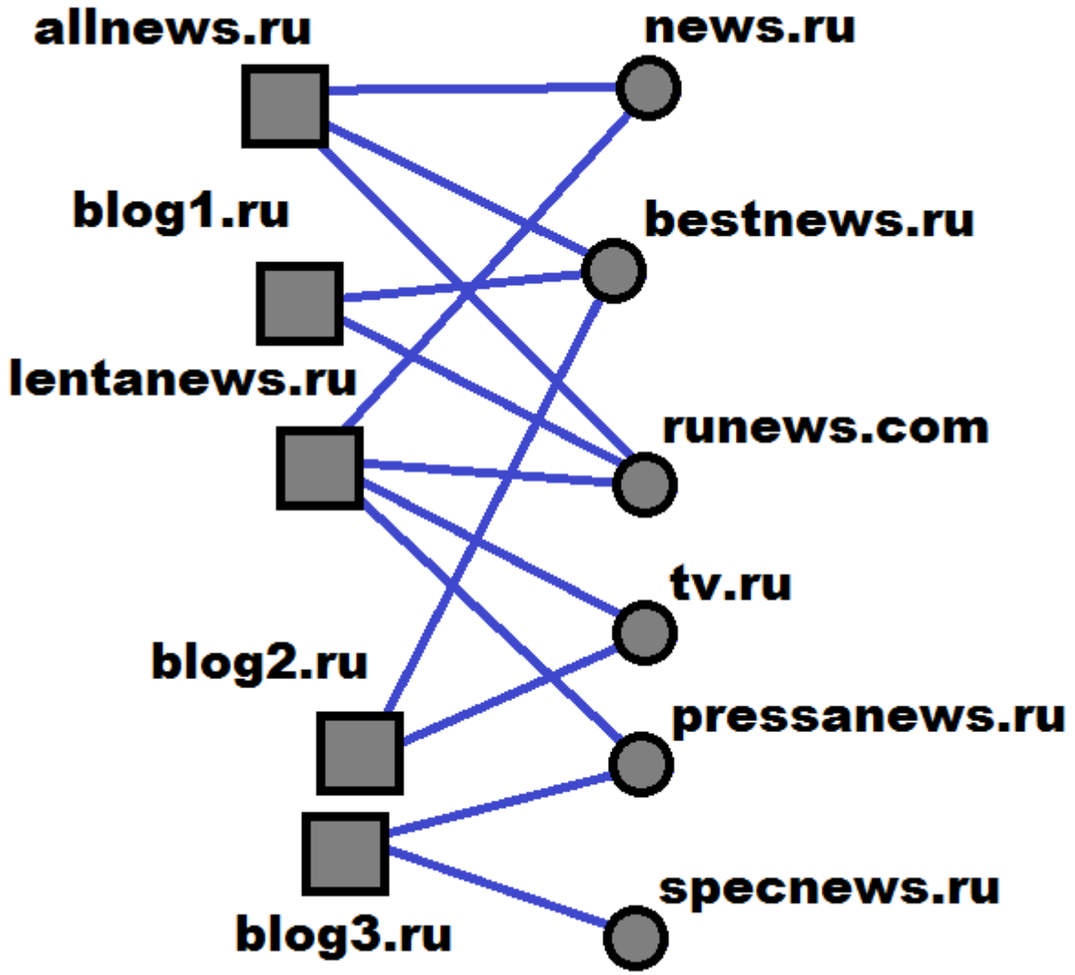
– ферма спама,

– спам

Для формирования доверенной зоны можно использовать эксперта

Ещё итерационные алгоритмы: HITS

Агрегаторы

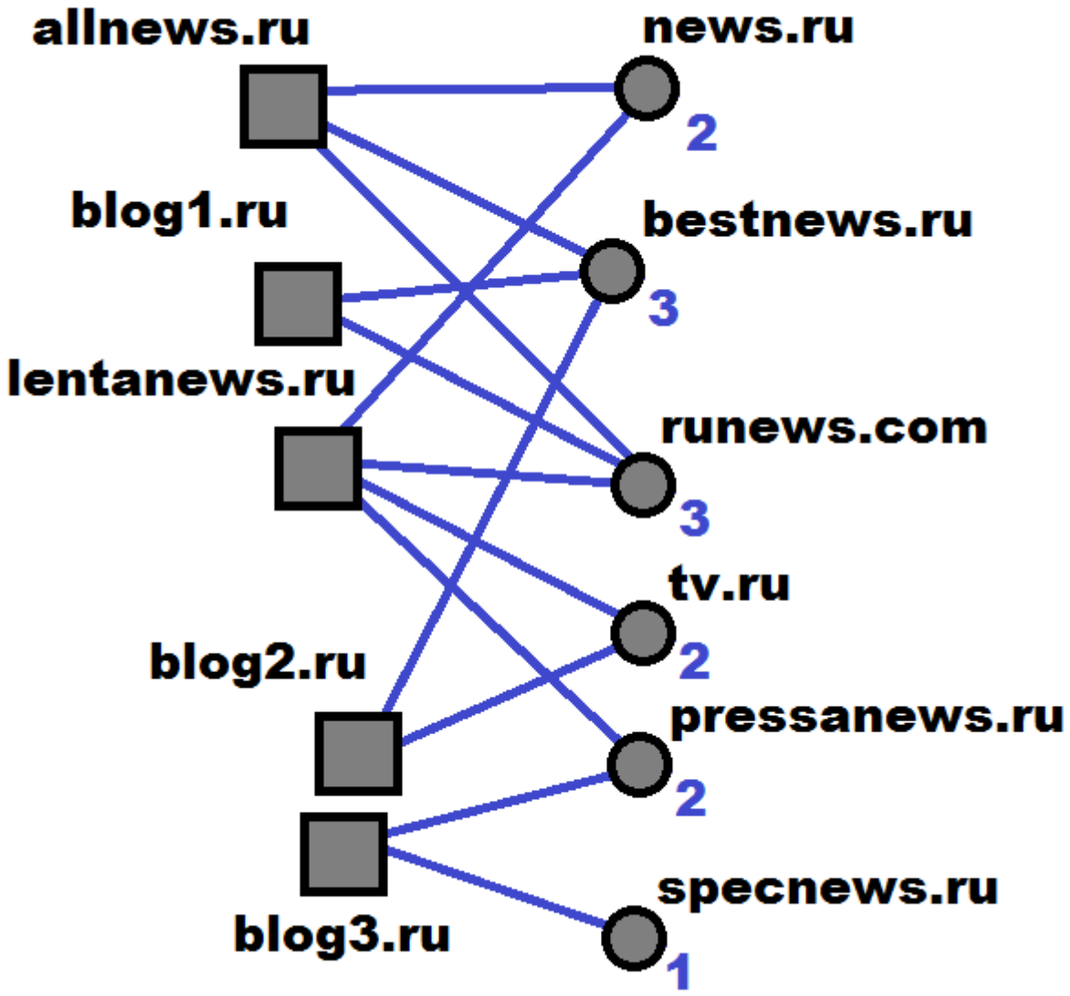


Новостные сайты

поиск ценных источников информации

Ещё итерационные алгоритмы: HITS

Агрегаторы

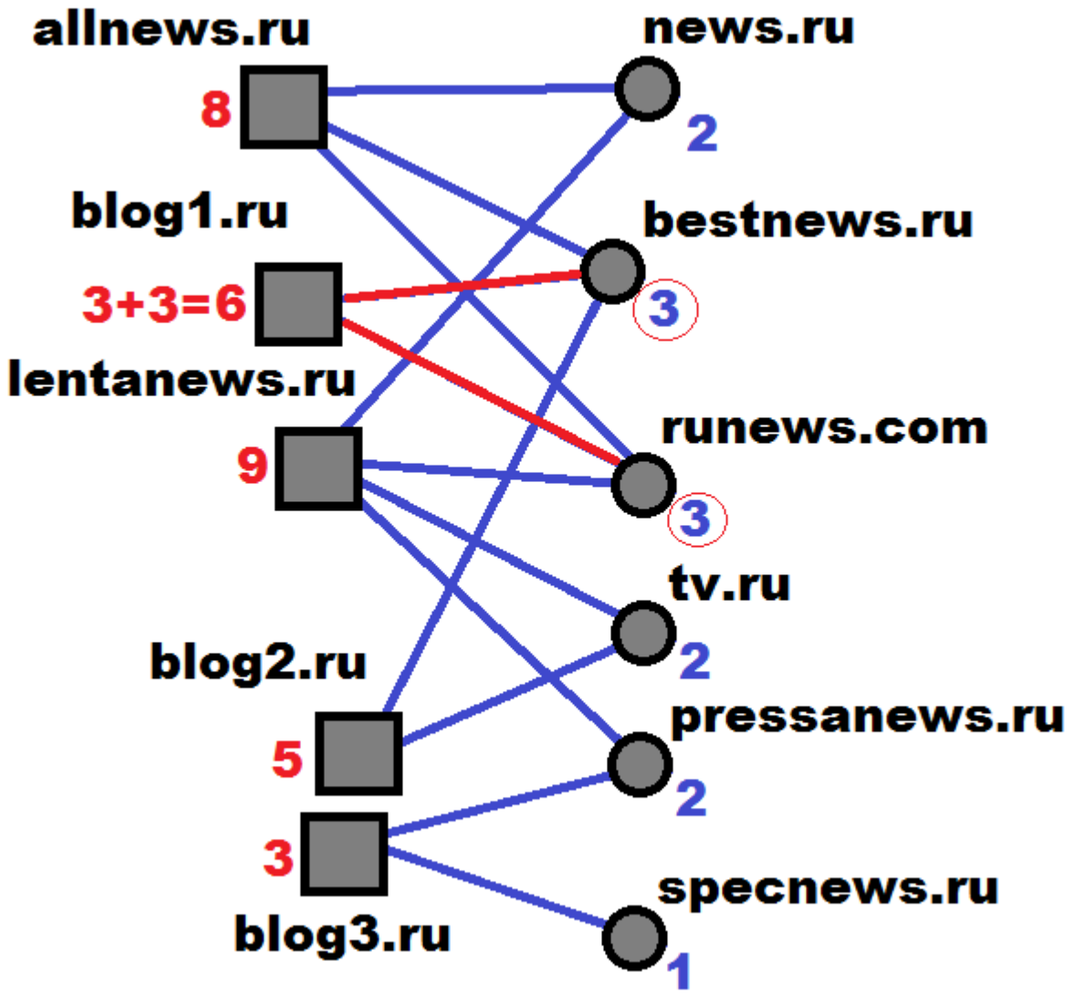


Новостные сайты

Ценное то – на что ссылаются

Ещё итерационные алгоритмы: HITS

Агрегаторы



Новостные сайты

Ценное то – на что ссылаются

Ещё итерационные алгоритмы: HITS

Дальше идея понятна…

К решению какого матричного уравнения всё сводится?

Какая задача здесь возникает?

http://liacs.leidenuniv.nl/~takesfw/SNACS/lecture4.pdf

HITS=«Hyperlink Induced Topic Search» (алгоритм Клейнберга)

Пусть в графе вершины $V = H \cup A$**:** $H = \{h_i\}, A = \{a_j\}$**,**

рёбра $E \subseteq H \times A$**,**

1. Инициализация:

$$w(h_i) = \frac{1}{|H|}, w(a_j) = \frac{1}{|A|}$$

2. Повторять

$$w(a_j) = \sum_{(i,j)\in E} w(h_i),\ w(h_i) = \sum_{(i,j)\in E} w(a_j)$$

$$w(a_j) = \frac{w(a_j)}{\sum_t w(a_t)},\ w(h_i) = \frac{w(h_i)}{\sum_t w(h_t)}$$

до сходимости

$$\sum_t w(h_t) < \varepsilon,\ \sum_t w(a_t) < \varepsilon$$

HITS

$$\begin{cases} a = M^{\mathrm{T}} h \\ h = Ma = MM^{\mathrm{T}} h \end{cases}$$

$$\begin{cases} a^{(t)} = M^{\mathrm{T}} h^{(t-1)} \\ h^{(t)} = MM^{\mathrm{T}} h^{(t-1)} = (MM^{\mathrm{T}})^t h^{(0)} \end{cases}$$

Иногда используют другие нормировки

Недостатки:

– Строгое разграничение: хаб / ресурс

– Надо нормировать, в отличие от PageRank

Kleinberg, Jon «Hubs, Authorities, and Communities» Cornell University. 1999.

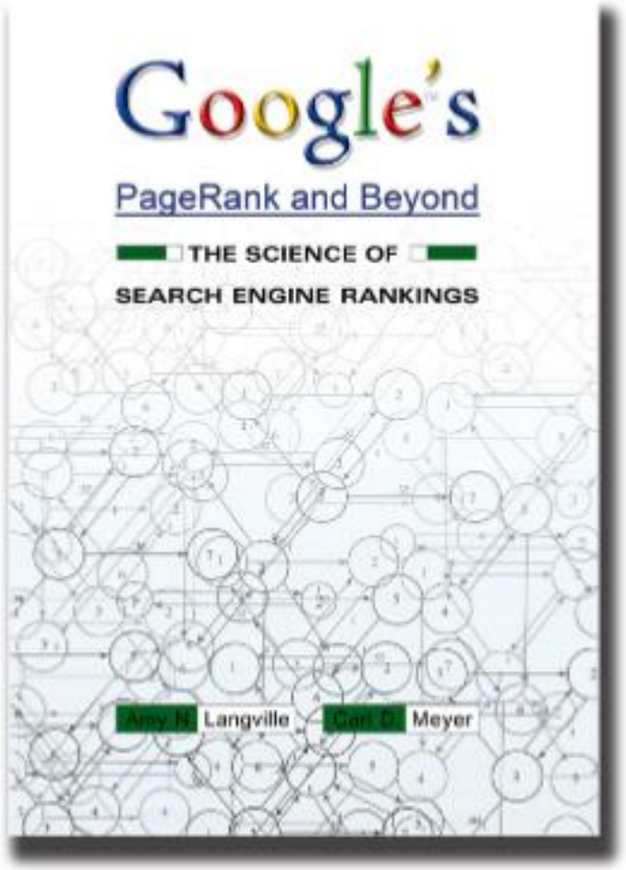
Case: где ещё применяется

«Impact Factor» научных журналов
– среднее число цитирований статей,
опубликованных в этом журнале за последние 2 года

«New Lung Cancer Study Takes Page from Google’s Playbook»
http://www.scripps.edu/news/press/2013/20130325lung_cancer.html

Что почитать

«Google's PageRank and beyond»

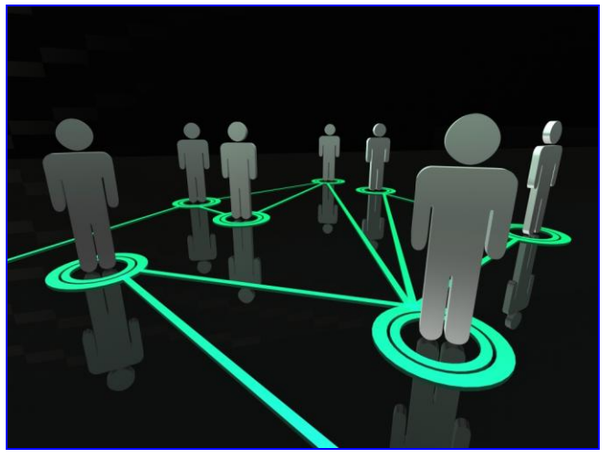


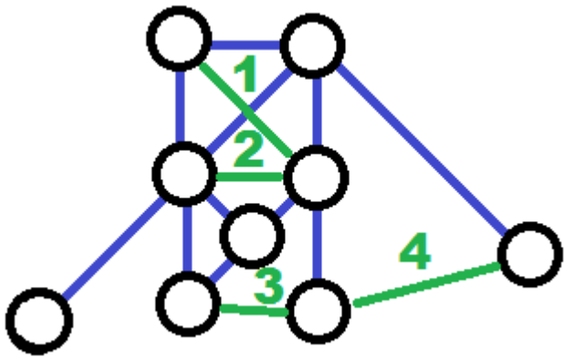
http://geza.kzoo.edu/~erdi/patent/langvillebook.pdf

case: Прогнозирование появления ребра в динамическом графе (Link Prediction Problem)

Международное соревнование «IJCNN Social Network Challenge»

http://www.kaggle.com/c/socialNetwork/



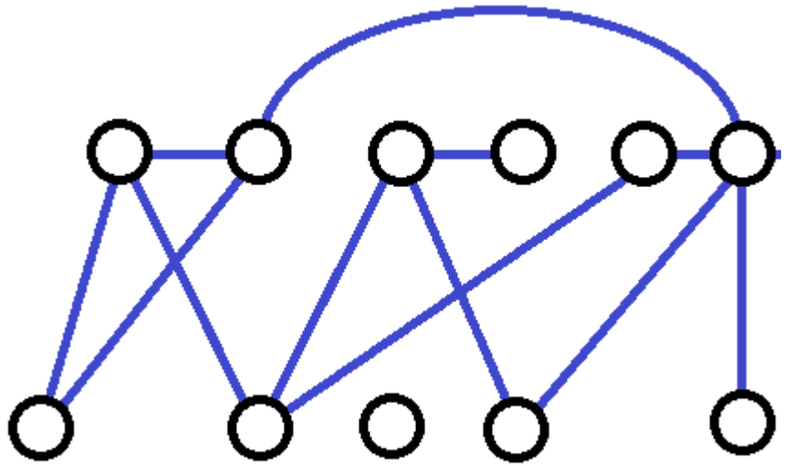


Дан граф,
Список потенциальных рёбер

Необходимо ранжировать список по вероятности появления

Соревнование «IJCNN Social Network Challenge»

Задача не в стандартной постановке –
граф почти двудольный, ориентированный!



вершин = 1’100’000

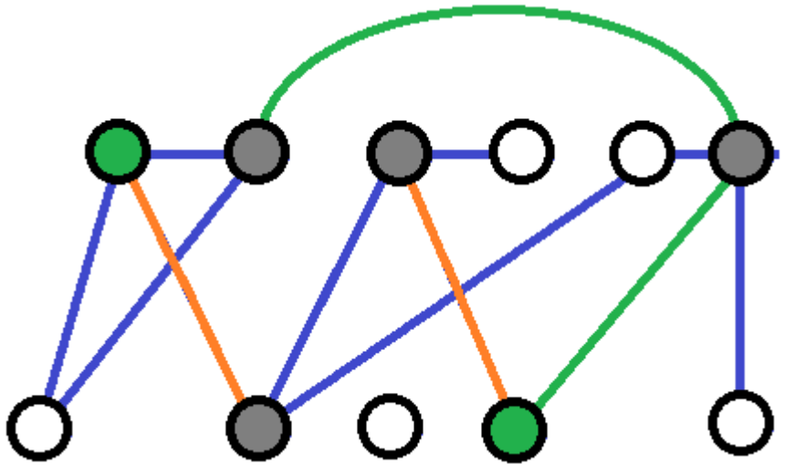
рёбер = 7’200’000

Сеть Flickr

Тест = 4480+4480
потенциальных рёбер

Как решать?

Описанные признаки легко обобщаются на двудольный случай



Кстати, тонкости в задаче –

как выбрать обучающую выборку (надо знать как делал заказчик)!

Если не-рёбра = случайные не рёбра,
то задача лёгкая, обобщения нет

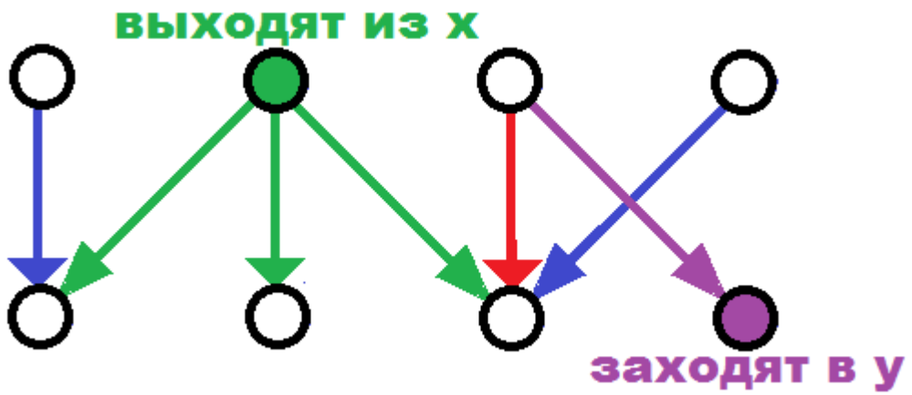
Если не-рёбра = почти рёбра,
то они могут скоро стать рёбрами… а этому мы и должны научиться

Первый подход

друг друга

$$\frac{|(\Gamma(x,*)\times\Gamma(*,y))\cap E|}{|\Gamma(x,*)|\cdot|\Gamma(*,y)|+1}$$

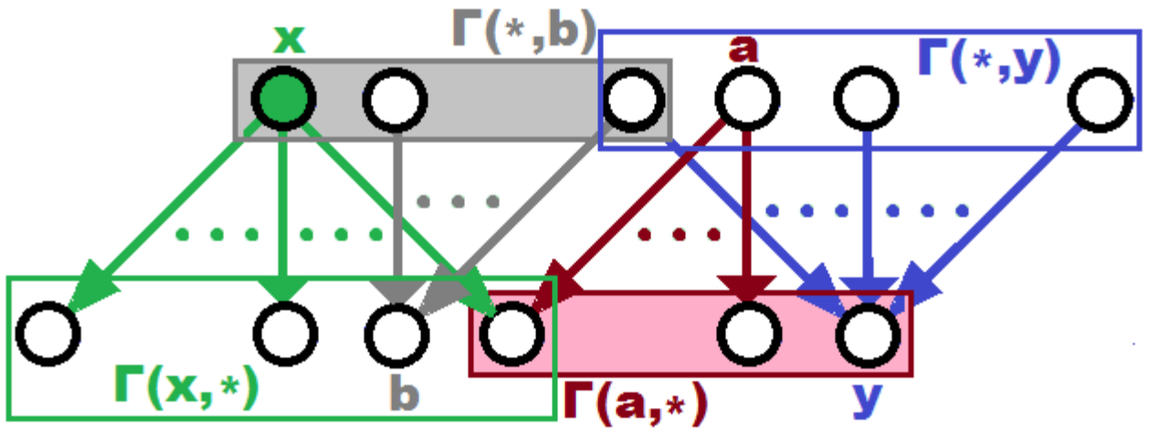
$$\Gamma(x,*)=\{y\in V\,|\,(x,y)\in E\}$$



Улучшение качества при таком признаке

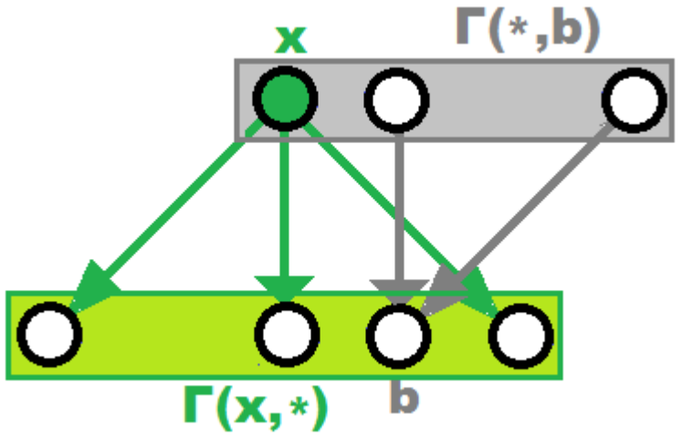
$$\sum_{\substack{a\in\Gamma(*,y)\\ b\in\Gamma(x,*)}} \frac{|\Gamma(a,*)\cap\Gamma(x,*)|\cdot|\Gamma(*,b)\cap\Gamma(*,y)|}{\sqrt{|\Gamma(a,*)|\cdot|\Gamma(*,b)|}} \Big/ |\Gamma(x,*)|\cdot|\Gamma(*,y)|+1$$

Какой смысл этого признака?



Признак №2

$$\frac{1}{|\Gamma(x,*)|}\sum_{b\in\Gamma(x,*)}\frac{|(\Gamma(*,b)\cap\Gamma(x,*))\cap E|}{|\Gamma(*,b)|\cdot|\Gamma(x,*)|+1}$$



насколько дружелюбны друзья x

(не зависит от y, хорош в комбинации)

Второй подход

вершины соединены, если соединены похожие

$$\frac{|(X \times Y) \cap E|}{|X| \cdot |Y| + 1}$$

X **– вершины похожие на** x**,**

Y **– вершины похожие на** y**.**

Что такое похожие?

сравниваем как строки в матрице смежности

Лучшее – скалярное произведение с довеском:

$$|\Gamma(x,*) \cap \Gamma(a,*)| - \frac{1}{2 + |\Gamma(a,*)| - |\Gamma(x,*) \cap \Gamma(a,*)|}$$

Оптимальные множества: $|X| = 9, |Y| = 40$

При разных метриках – некоррелированные признаки

Как учитывать похожесть?

Вместо $\frac{|(X \times Y) \cap E|}{|X| \cdot |Y| + 1}$

весовую схему

$$\frac{1}{|X| \cdot |Y| + 1} \sum_{a \in A} \sum_{b \in B} w(a) w'(b)$$

Блендинг

I = 87.5

I + II = 90.7

III = 90.7

I + II + III = 92.6

PR = 93.0

I + II + III + PR = 95.0

Итог

Есть много способ генерации признаков

В классике «важность» – центральность

Очень хорошие признаки – на случайных блужданиях

Можно модифицировать блуждания

(получаем много применений)