Simrank++

Query Rewriting through Link Analysis of Click Graph

Alekh Jindal

Motivation

Google | investment | Search | Advanced Search | Preferences

Web Results **1** - **10** of about **270,000,000** for **investment** [d

[450%+ **Investment** Returns](#) Sponsored Link
www.seismaresearch.com High Yield Alternative **Investments**. We Show You How. Start Here.

Related searches: investment **options** investment **definition** **real estate** investment **stocks**

Investment - Wikipedia, the free encyclopedia
Investment or investing [1] is a term with several closely-related meanings in business management, finance and economics, related to saving or deferring **...**
en.wikipedia.org/wiki/**Investment** - 47k - Cached - Similar pages

Union **Investment** - Ihr professioneller Partner im Asset Management - [Translate this page]
Ergänzen Sie Ihre Rente damit Sie später gut versorgt sind! Die Union **Investment** Gruppe ist einer der größten deutschen Asset Manager für private und **...**
www.union-**investment**.de/ - 12k - Cached - Similar pages

DWS **Investments** International :: HOME
Investment trends, insights and analysis on the key market developments brought **...** DWS **Investments**, a member of Deutsche Bank Group is one of the world's **...**
www.dws.com/ - 28k - Cached - Similar pages

Germany Trade and Invest - Home
Germany Trade and Invest is the new foreign trade and inward **investment** agency **...** We inform you about **investment** opportunities in Germany and the general **...**
www.invest-in-germany.com/ - 12k - Cached - Similar pages

www.bund.de **Investment**
The official business portal of the state of Hesse provides extensive information on the business location Hesse, conditions for **investment** and further **...**
www.bund.de/nn_262504/Fremdsprachen/Struktur/EN/Economy-and-Trade/Economy-and-Trade-knoten.html__nnn=true - 28k - Cached - Similar pages

Buckhead **Investment** Partners - Home
Buckhead **Investment** Partners is a Registered **Investment** Advisor that offers a wide array of **investment** strategies and services. **...**
www.buckhead**investments**.com/ - 12k - Cached - Similar pages

Sponsored Li

Make 12% per mo. in ETFs
Proven system. Trade only 5 minu
per day. Easier than 4x or options
www.ETFtradingcourse.com

Top China-Fonds
5 Sterne von Morningstar.
Vom Wachstum in China profitiere
www.Fidelity.de

Forex **Investments**
Free $100,000 Practice Account W
Real-Time Charts, News & Resea
www.ac-markets.com

Smart Stock Finance
Earn 250 % For 30 Days
Payment By LibertyReserve.
www.smartstockfinance.com

$2500/Day Guaranteed
Make Money 100% on Autopilot !
Step-by-Step Instructions - $4.95
www.ForexAutoMoney.com

Making Money Transfer
Have The Chance To Earn
The High Profit From Forex Marke
www.makingmoneytransfer.com

Property Deal Analysis
Assess your **investment** returns,
improve your performance. Free!
www.PropAble.com

Motivation

Google profitable investment Search Advanced Search Preferences

Web Results **1** - **10** of about **631,000** for profitable investment. (**0.23** seconds)

Today's Most **Profitable Investments** |
23 Mar 2006 **...** In this issue, Dr. Mark Skousen shares his two most **profitable investments** and wealth-building strategies. His answers may surprise you.
www.**investment**u.com/IUEL/2006/20060323.html - 37k - Cached - Similar pages

Aegypten issue tracker: Issue 1086: MONEY AND GOLD DUST FOR **...**
Title, MONEY AND GOLD DUST FOR **PROFITABLE INVESTMENT**. Priority, Status, unread. Superseder, Nosy List, troshev.troshev, troshev.troshev1 **...**
www.intevation.de/roundup/aegypten/issue1086 - 13k - Cached - Similar pages

Pakistan- A **Profitable Investment** Destination
Pakistan Embassy Berlin: Visa application form, Passport application form.
www.pakemb.de/index.php?id=128 - 10k - Cached - Similar pages

dict.cc dictionary :: **profitable investment** :: English-German **...**
dict.cc English-German Dictionary: Translation for **profitable investment**.
www.dict.cc/english-german/**profitable+investment**.html - 9k - Cached - Similar pages

Profitable Investment Portfolio
A number of diplomats and Ambassadors have expressed the keen desire of their respective governments to explore opportunities of **profitable investments** in ... **...**
www.**profitableinvestment**portfolio.com/ - 83k - Cached - Similar pages

Learn how to Invest, Buy Stocks, Sell Stocks, Investing and **...**
To bring you in a position to accomplish **profitable investments** in equities, easily manage to make money now and plan for your financial future and **...**
www.greekshares.com/ - 44k - Cached - Similar pages

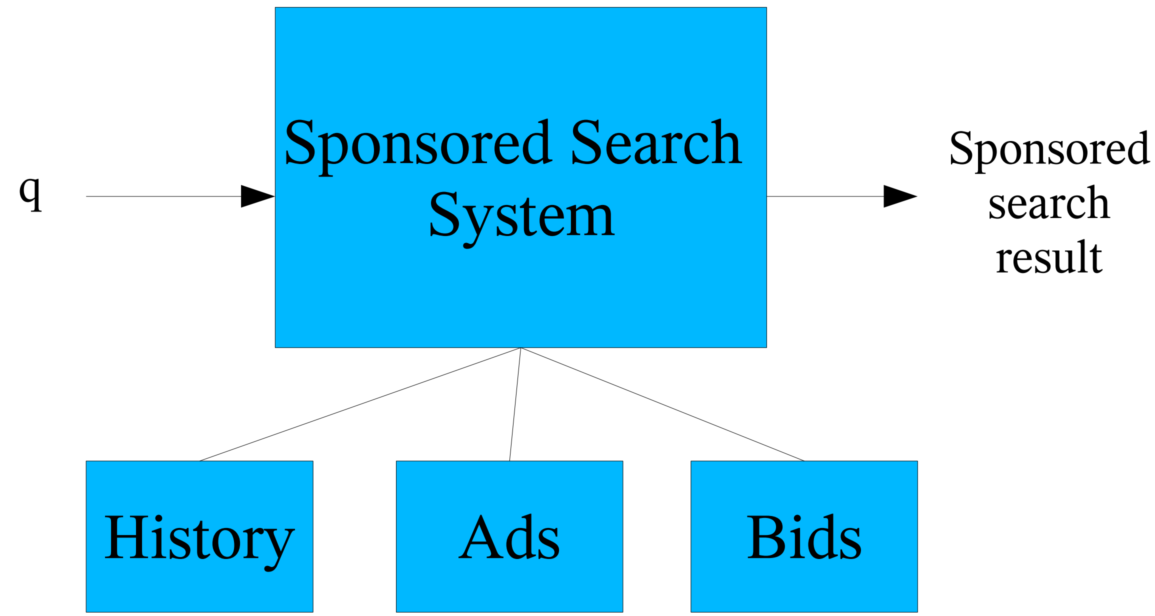
Schneider Electric - Training - Your benefits - Training, a **...**
Training, a **profitable investment**. Return on **investment**: immediate and over time. •

Sponsored Search

- Ads relevant to user query shown above or alongside search results
- Each bid has query(q), ad(α) and price(p)

Sponsored Search

- Ads relevant to user query shown above or alongside search results
- Each bid has query(q), ad(α) and price(p)

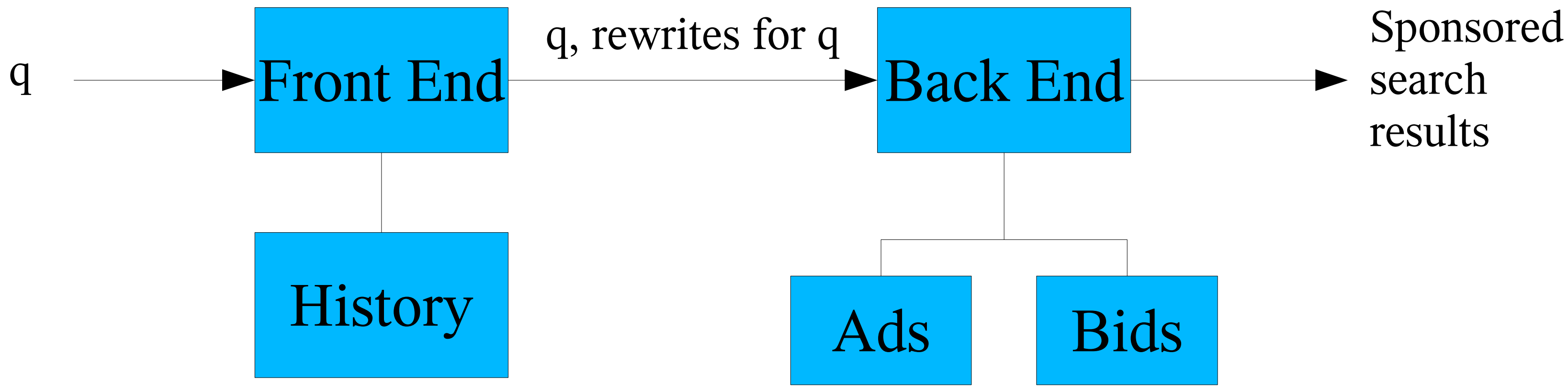


Query Rewrites

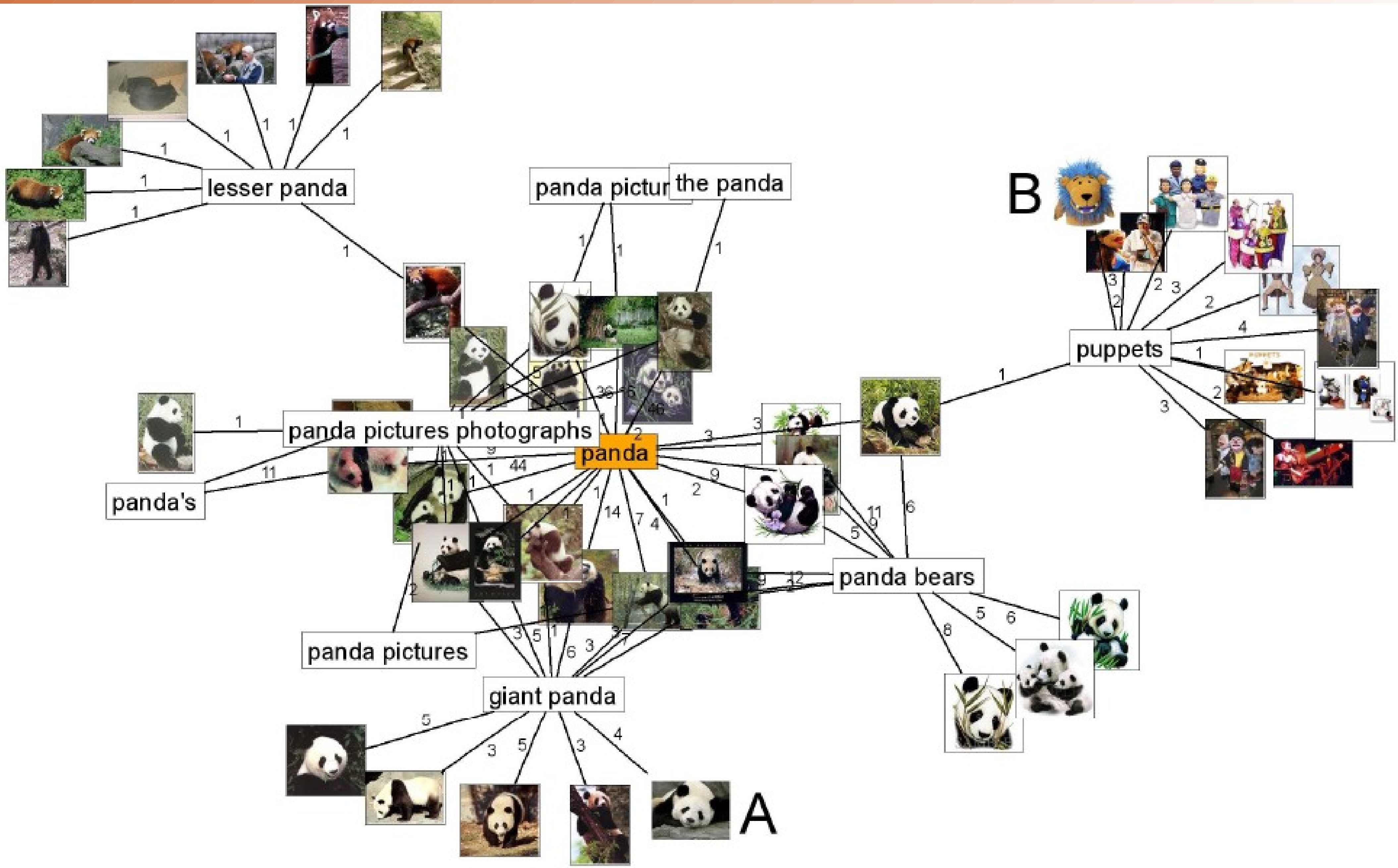
- Not direct bids for many queries
- Ads have little text; lesser information
- Rewrites: similar queries based on history of ads displayed and clicked

Query Rewrites

- Not direct bids for many queries
- Ads have little text; less information
- Rewrites: similar queries based on history of ads displayed and clicked



Click Graph

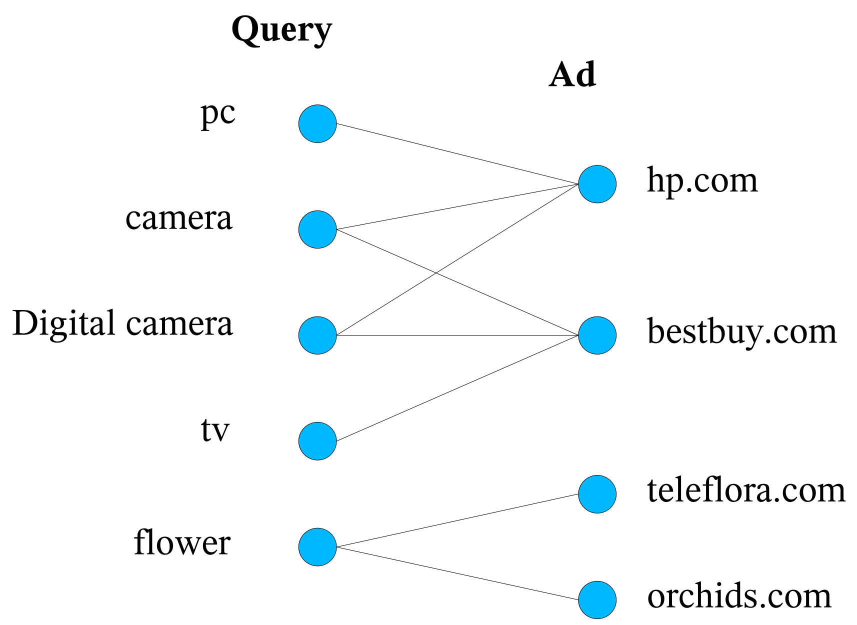


lesser panda
panda pictur
the panda
B
puppets
panda pictures photographs
panda
panda's
panda bears
panda pictures
giant panda
A

Click Graph

- Generated by back end
- Directed, weighted, bipartite graph
- Formally: $G = (Q, A, E)$
- Q: set of queries q
- A: set of ads α
- E: set of edges e from q to α, s.t. at least one user that issued q clicked on α
- Edge weights:
  - Impressions
  - Clicks
  - Expected click rate

Click Graph - example

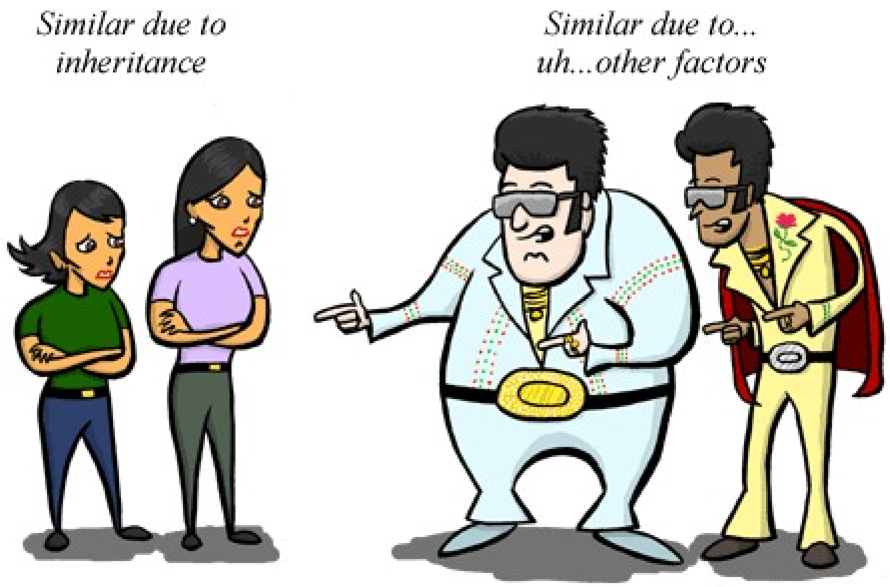


Query Similarity

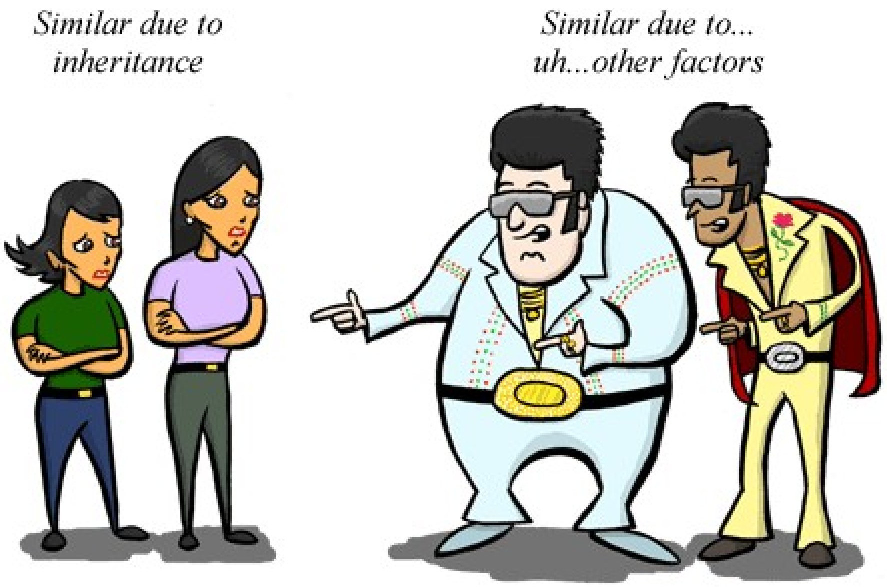
- Goal: find similar queries
- Intuition: queries with common ad clicks are similar
- Analogous to collaborative filtering (CF)
  - Users as queries; Recommendation as ads

Query Similarity

Query Similarity

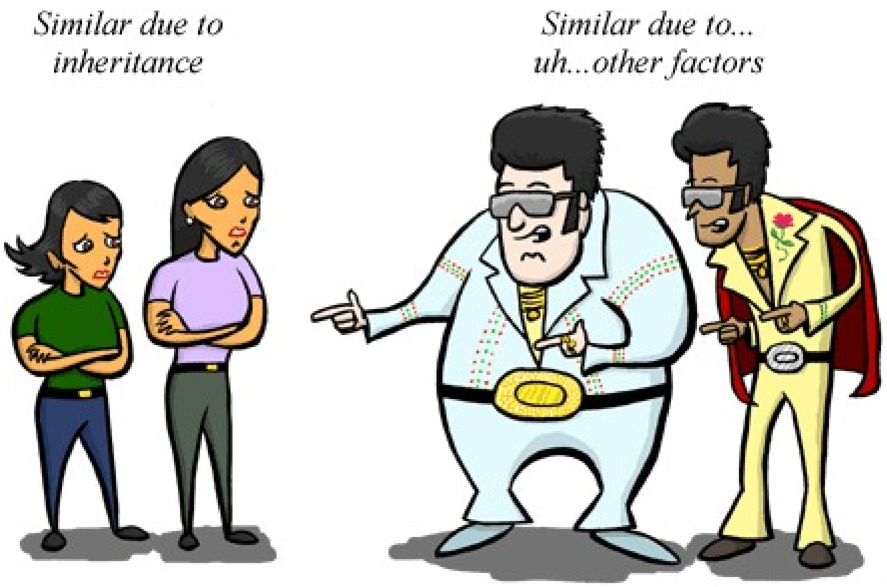


Query Similarity



Guys are similar if they like the similar girl!

Query Similarity



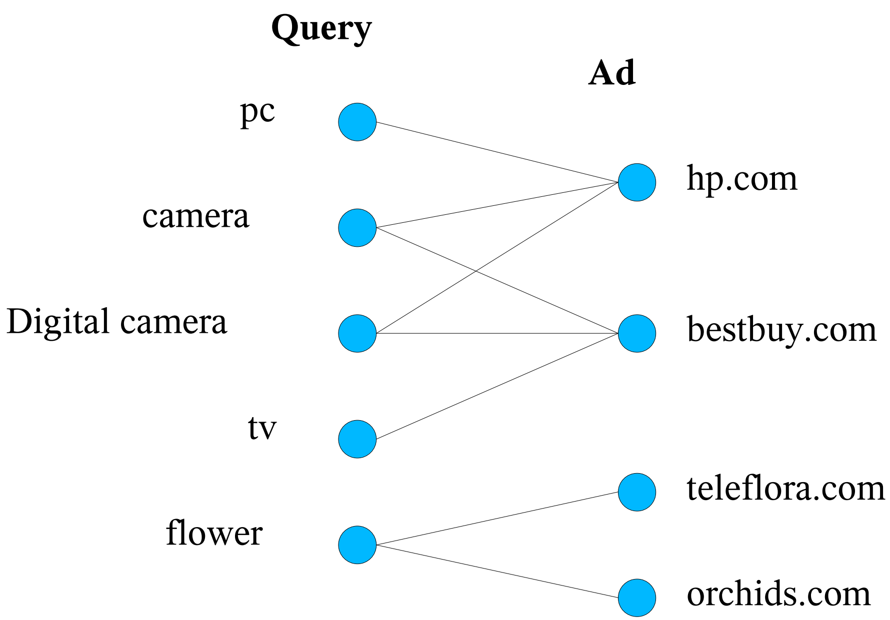
Guys are similar if they like the similar girl!
... and vice-versa!

Similarity: naïve approach

Idea: count number of common ads

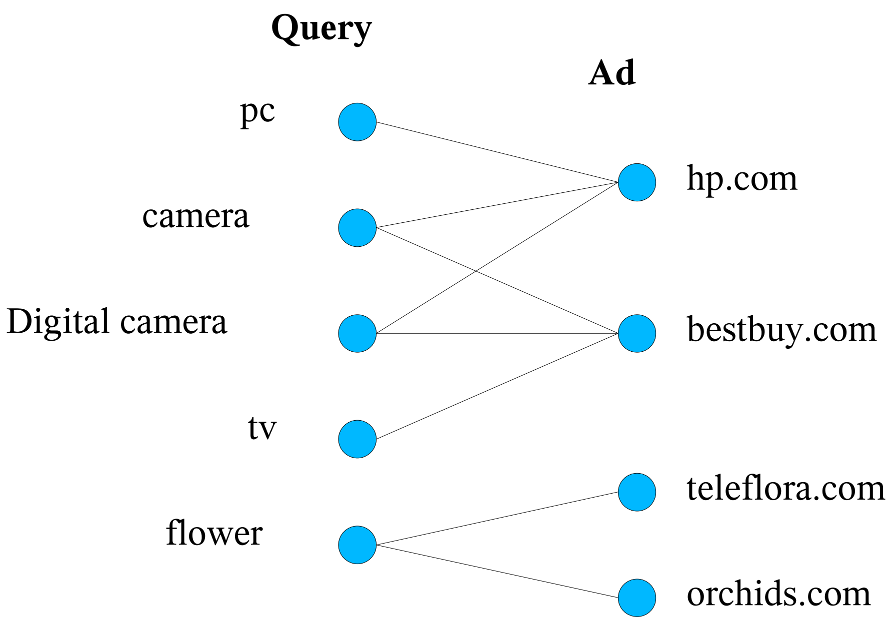
Similarity: naïve approach

Idea: count number of common ads



Similarity: naïve approach

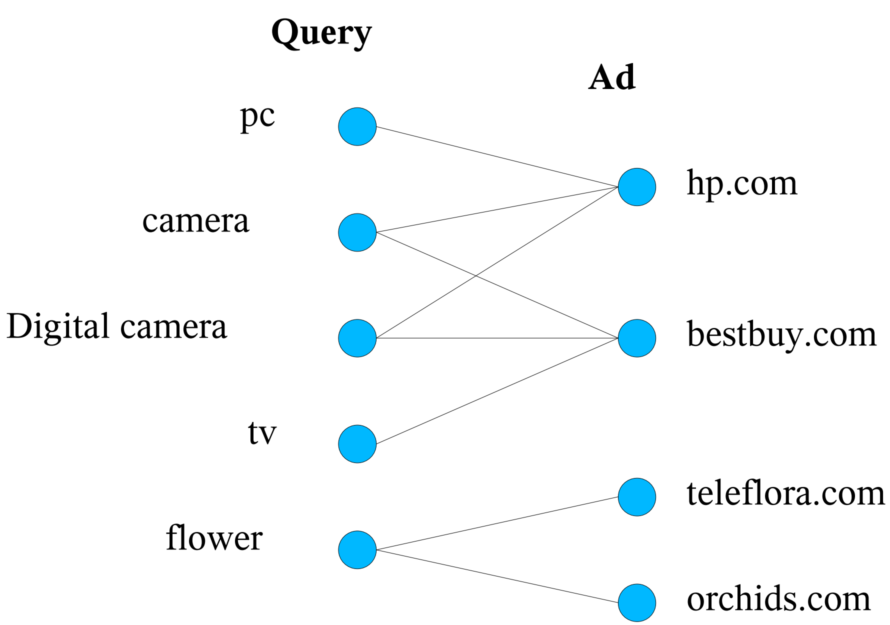
Idea: count number of common ads



	pc	camera	digital camera	tv	flower
pc	-	1	1	0	0
camera	1	-	2	1	0
digital camera	1	2	-	1	0
tv	0	1	1	-	0
flower	0	0	0	0	-

Similarity: naïve approach

Idea: count number of common ads



	pc	camera	digital camera	tv	flower
pc	-	1	1	0	0
camera	1	-	2	1	0
digital camera	1	2	-	1	0
tv	0	1	1	-	0
flower	0	0	0	0	-

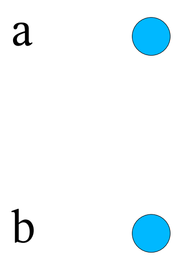
Problem: sim(pc, tv) = 0

Similarity: Simrank

Idea: Two objects are similar if they are referenced by similar objects

Similarity: Simrank

Idea: Two objects are similar if they are referenced by similar objects

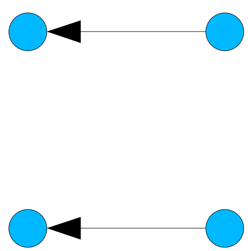


Similarity: Simrank

Idea: Two objects are similar if they are referenced by similar objects

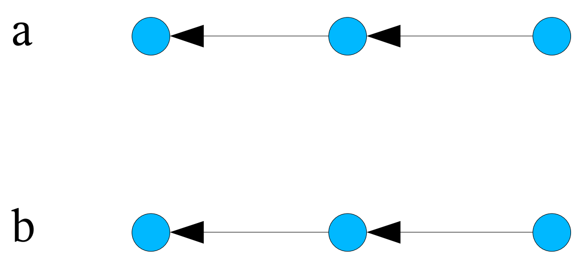
a

b



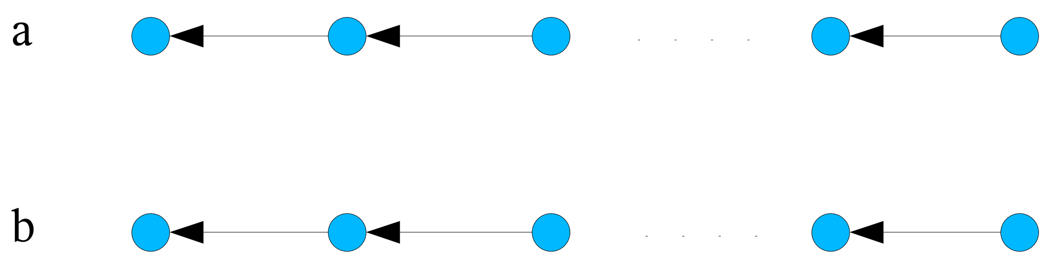
Similarity: Simrank

Idea: Two objects are similar if they are referenced by similar objects



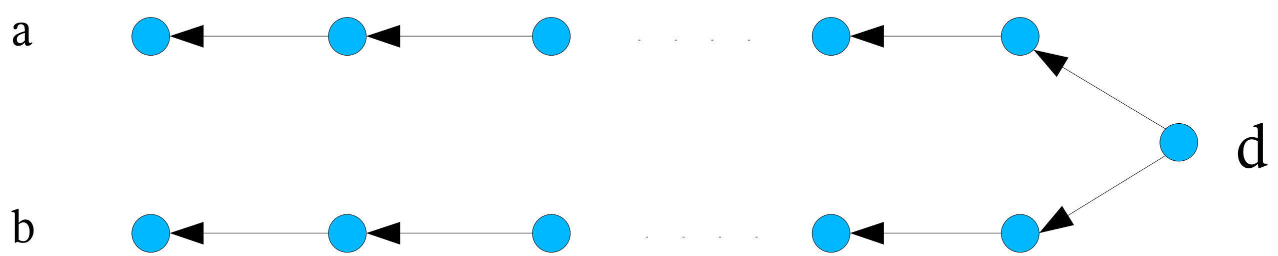
Similarity: Simrank

Idea: Two objects are similar if they are referenced by similar objects



Similarity: Simrank

Idea: Two objects are similar if they are referenced by similar objects

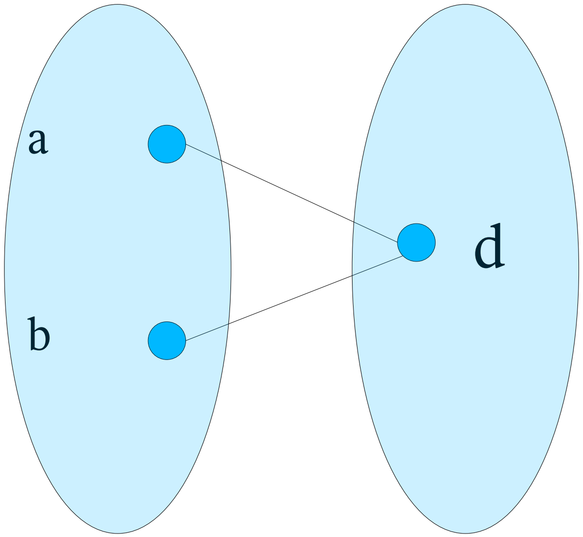


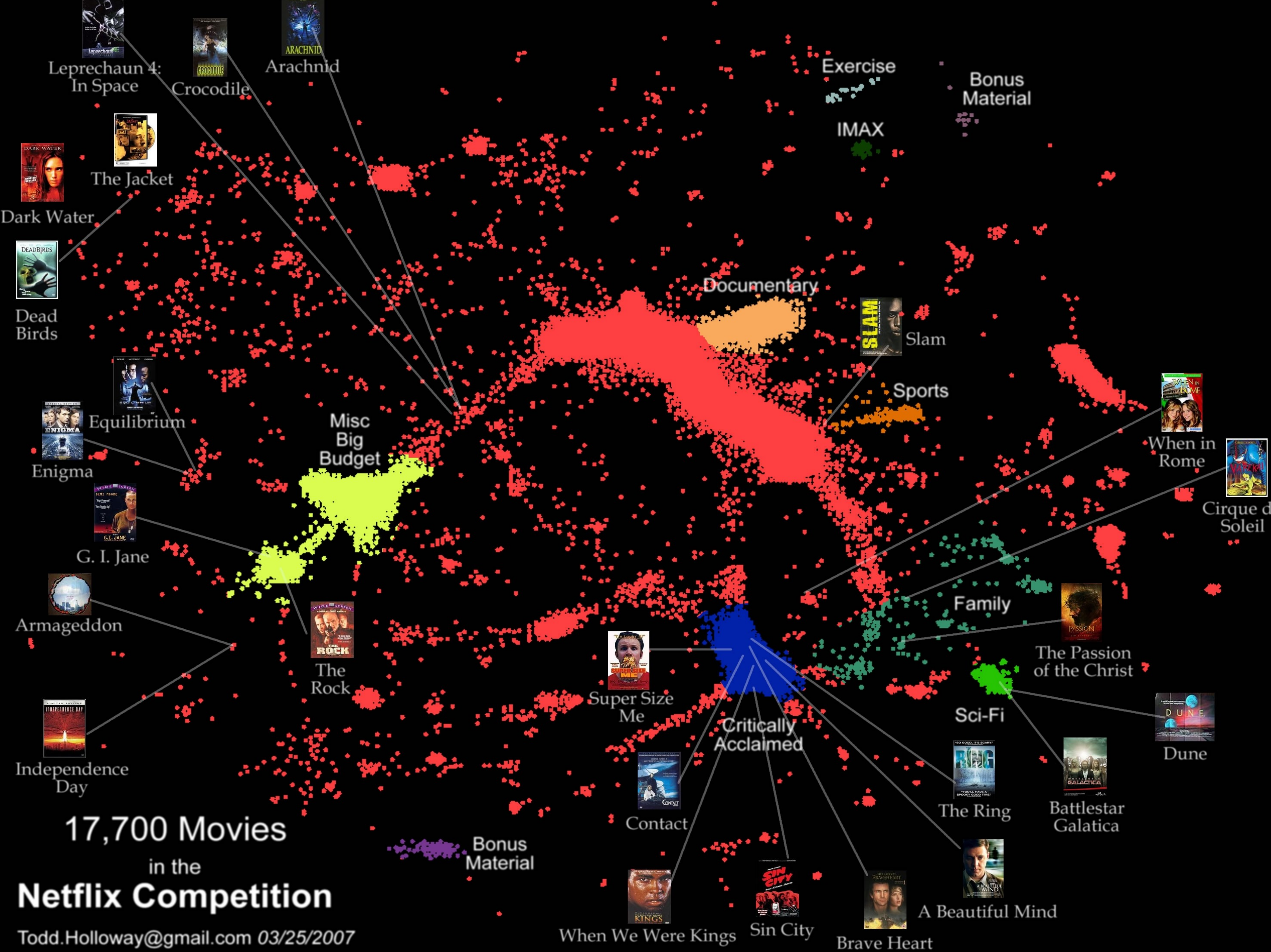
Similarity: Bipartite Simrank

Idea: Two objects of one type are similar if they are referenced by similar objects of second type

Similarity: Bipartite Simrank

Idea: Two objects of one type are similar if they are referenced by similar objects of second type




Leprechaun 4: In Space
Crocodile
Arachnid
Exercise
Bonus Material
IMAX
The Jacket
Dark Water
Dead Birds
Documentary
Slam
Sports
Equilibrium
Enigma
Misc Big Budget
When in Rome
Cirque d Soleil
G. I. Jane
Family
Armageddon
The Rock
Super Size Me
The Passion of the Christ
Sci-Fi
Critically Acclaimed
Dune
Independence Day
Contact
The Ring
Battlestar Galatica
17,700 Movies
in the
Netflix Competition
Bonus Material
A Beautiful Mind
When We Were Kings
Sin City
Brave Heart
Todd.Holloway@gmail.com 03/25/2007

Similarity: Bipartite Simrank

- Formally: E(x) is the set of neighbors of x
- N(x) is the number of neighbors of x
- For queries q and q', similarity s(q,q') is given as:

$$s(q,q') = \frac{C_1}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} s(i,j)$$

- Similarly, for ads α and α', similarity s(α,α') is given as:

$$s(\alpha,\alpha') = \frac{C_2}{N(\alpha)N(\alpha')} \sum_{i \in E(\alpha)} \sum_{j \in E(\alpha')} s(i,j)$$

- C_1, C_2 are constants in [0,1]

Similarity: Bipartite Simrank

- Example: Find s(a,c)

Let $C_1 = 0.8$

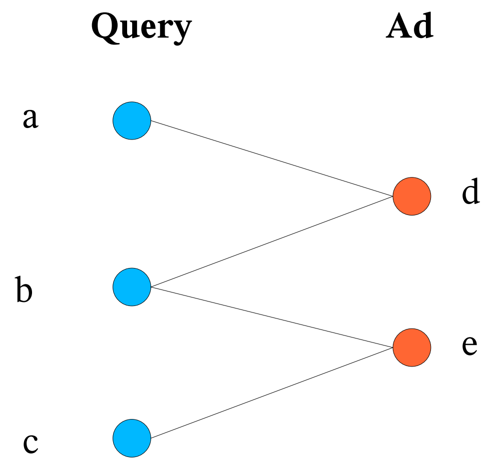
$$s(a,c) = \frac{C_1}{N(a)N(c)} \sum_{i \in E(a)} \sum_{j \in E(c)} s(i,j)$$

$$s(a,c) = 0.8 \,.\, s(d,e)$$

Iteration 1:

s(x,x) = 1, s(x,y) = 0

s(a,c) = 0



Similarity: Bipartite Simrank

- Iteration 2:

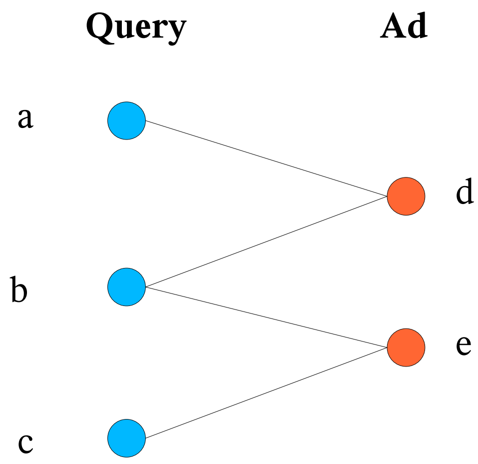
$$s(a,c) = 0.8\ .\ s(d,e)$$

$$s(a,c) = 0.8\ .\ \frac{C_2}{N(d)N(e)} \sum_{i \in E(d)} \sum_{j \in E(e)} s(i,j)$$

$$s(a,c) = 0.8\ .\ \{\frac{0.8}{2 \times 2}$$
$$(s(a,b) + s(a,c) + s(b,b) + s(b,c))\}$$

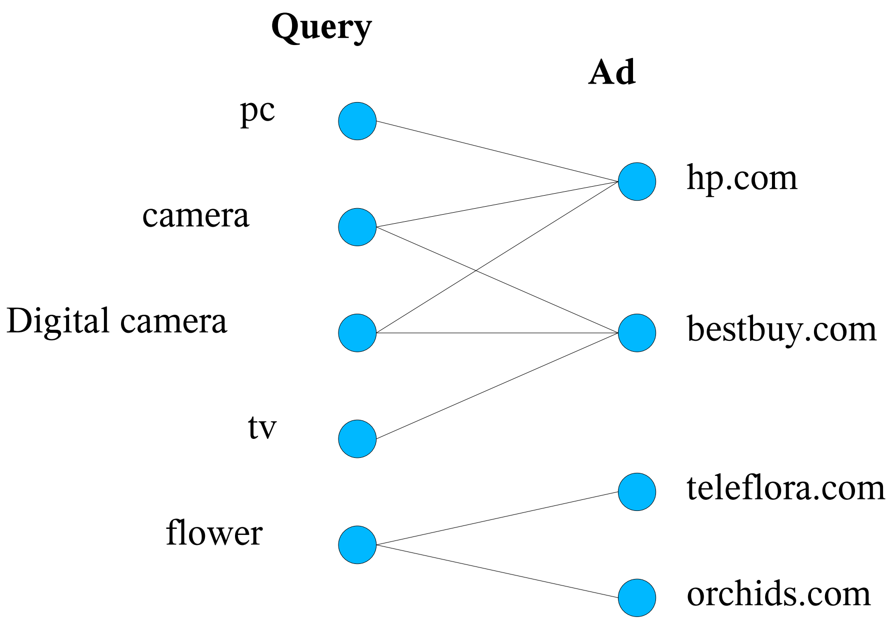
$$s(a,c) = 0.32\ .\ \{(0+0+1+0)\}$$

$$s(a,c) = 0.32$$



Similarity: Bipartite Simrank

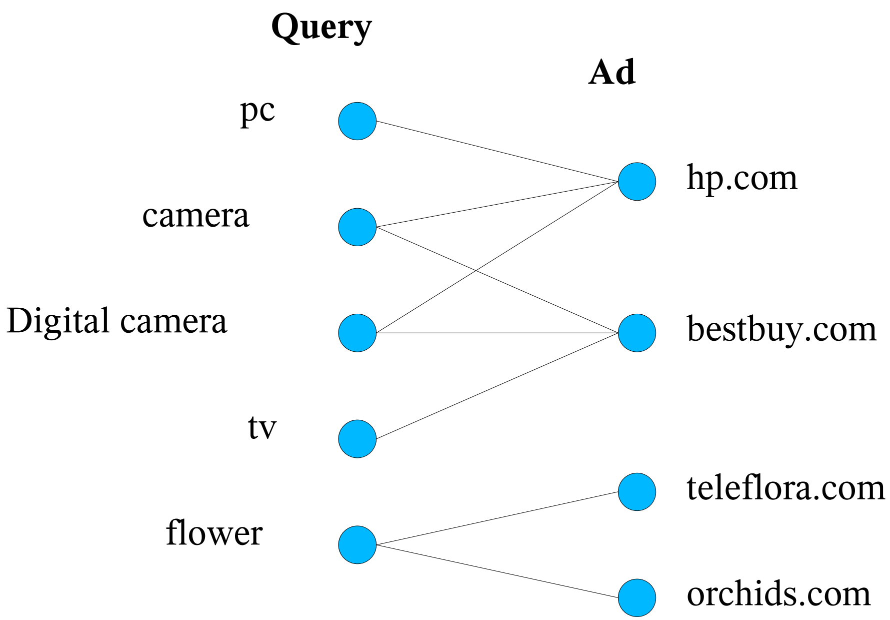
Idea: Two objects of one type are similar if they are referenced by similar objects of second type



	pc	camera	digital camera	tv	flower
pc	-	0.619	0.619	0.437	0.000
camera	0.619	-	0.619	0.619	0.000
digital camera	0.619	0.619	-	0.619	0.000
tv	0.437	0.619	0.619	-	0.000
flower	0.000	0.000	0.000	0.000	-

Similarity: Bipartite Simrank

Idea: Two objects of one type are similar if they are referenced by similar objects of second type

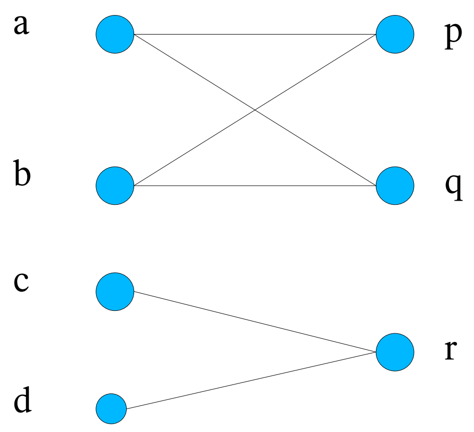


	pc	camera	digital camera	tv	flower
pc	-	0.619	0.619	0.437	0.000
camera	0.619	-	0.619	0.619	0.000
digital camera	0.619	0.619	-	0.619	0.000
tv	0.437	0.619	0.619	-	0.000
flower	0.000	0.000	0.000	0.000	-

s(camera,tv) = s(camera, digital camera)

Similarity: Bipartite Simrank

- “evidence” not taken into account



Iteration	sim(a,b)	sim(c,d)
1	0.4	0.8
2	0.56	0.8
3	0.624	0.8
4	0.6496	0.8
5	0.65984	0.8
6	0.663936	0.8
7	0.6655744	0.8

- sim(a,b) < sim(c,d)
- Expected: sim(a,b) > sim(c,d)

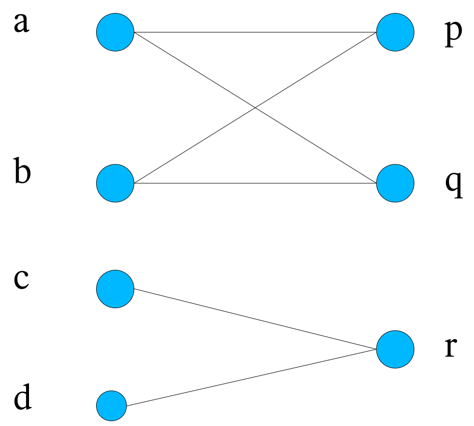
Similarity: Evidence Simrank

- “evidence”: Number of common neighbours
- Evidence function:

$$evidence(a,b) = \sum_{i=1}^{|E(a) \cap E(b)|} \frac{1}{2^i}$$

- Revised Simrank:
  - $s_{evidence}(q,q') = evidence(q,q')\ .\ s(q,q')$
  - $S_{evidence}(\alpha,\alpha') = evidence(\alpha,\alpha')\ .\ s(\alpha,\alpha')$

Similarity: Evidence Simrank

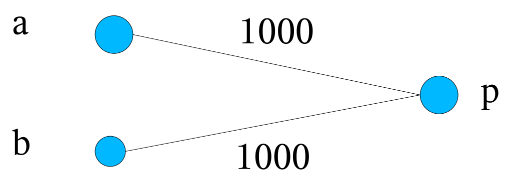


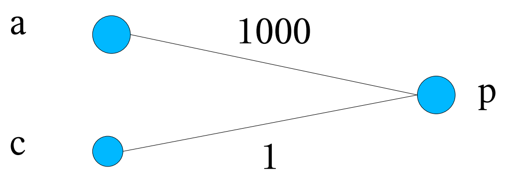
Iteration	sim(a,b)	sim(c,d)
1	0.3	0.4
2	0.42	0.4
3	0.468	0.4
4	0.4872	0.4
5	0.49488	0.4
6	0.497952	0.4
7	0.4991808	0.4

- sim(a,b) > sim(c,d) after 1st iteration

Similarity: Weighted Simrank

- Consistency Rules
  - If variance is less and edge weight more then simi-larity is more





  - Expected: sim(a,b) > sim(a,c)

Similarity: Weighted Simrank

- Consistency Rules
  - For equal variance, if edge weight is more then similarity is more

  - Expected: sim(a,b) > sim(a,c)

Similarity: Weighted Simrank

- Transition probability:

$$p(x, i) = \text{spread}(i)\ .\ \text{normalized\_weight}(x, i) = W(x, i)$$

$$\text{spread}(i) = e^{-\text{variance}(i)}$$

$$normalized\_weight(\alpha, i) = \frac{w(\alpha, i)}{\sum_{j \in E(\alpha)} w(\alpha, j)}$$

- Revised Simrank:

$$s_{weighted}(q, q') = evidence(q, q')\ .\ C_1\ .\ \sum_{i \in E(q)} \sum_{j \in E(q')} W(q, i) W(q', j) s_{weighted}(i, j)$$

$$s_{weighted}(\alpha, \alpha') = evidence(\alpha, \alpha')\ .\ C_2\ .\ \sum_{i \in E(\alpha)} \sum_{j \in E(\alpha')} W(\alpha, i) W(\alpha', j) s_{weighted}(i, j)$$

Scalability

- Query rewrites offline and in batch
- Space required: $O(N^2)$
  - N: total number of nodes (query+ad)
- Time Required: $O(kN^3)$
  - k: number of iterations
  - typical value, k=7
- Time complexity can be reduced to: $O(kN^2d)$
  - d: average of N(a).N(b)
  - d does not grow with N
- For 15 million queries, 14 million ads and 28 million edges, Simrank++ completes in 6 hours on a single machine

Experiments: baselines

- Three query rewriting techniques
  - Pearson:

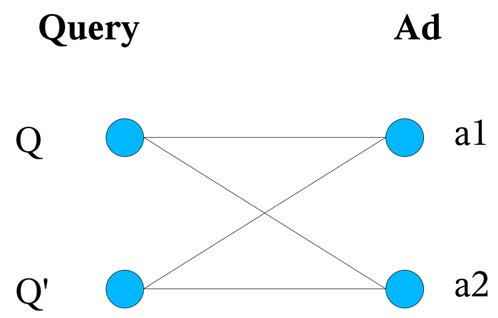
$$\mathrm{sim}_{pearson}(q,q') = \frac{\sum_{\alpha \in E(q) \cap E(q')} (w(q,\alpha) - \bar{w}_q)(w(q',\alpha) - \bar{w}_{q'})}{\sqrt{\sum_{\alpha \in E(q) \cap E(q')} (w(q,\alpha) - \bar{w}_q)^2 (w(q',\alpha) - \bar{w}_{q'})^2}}$$

  - Jaccard: $$\mathrm{sim}_{Jaccard}(q,q') = \frac{|E(q) \cap E(q')|}{|E(q) \cup E(q')|}$$

  - Cosine: $$\mathrm{sim}_{cosine}(q,q') = \arccos \frac{v(q) \,.\, v(q')}{\|v(q)\| \, \|v(q')\|}$$

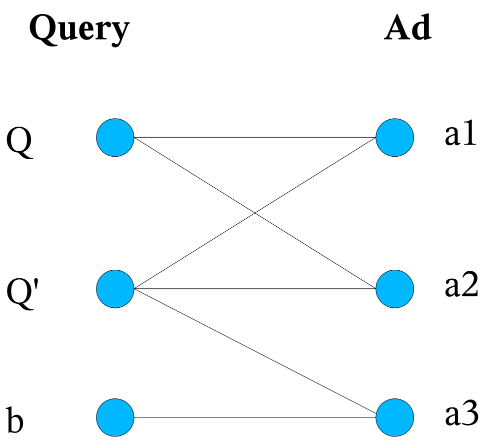
Experiments: baselines

- Example
  - $sim_{pearson}(Q,Q') = 1.414$
  - $sim_{Jaccard}(Q,Q') = 1.000$
  - $sim_{cosine}(Q,Q') = 0.841$



Experiments: baselines

- Example
  - $\text{sim}_{\text{pearson}}(Q,Q') = 1.414$
  - $\text{sim}_{\text{Jaccard}}(Q,Q') = 1.000$
  - $\text{sim}_{\text{cosine}}(Q,Q') = 0.841$



- However,
  - $\text{sim}_{\text{pearson}}(Q,b) = 0$
  - $\text{sim}_{\text{Jaccard}}(Q,b) = 0$
  - $\text{sim}_{\text{cosine}}(Q,b) = 0$

Experiments: Dataset

- Two week click graph from US Yahoo! Search
  - 15 million queries, 14 million ads, 28 million edges
- Edge weight: expected click rate
- Dataset partitioned into 5 big enough subgraphs
- Query set
  - Sampled from the same two-week period
  - Filter out the ones not present in subgraphs
  - 120 such queries

Experiments: Metrics

- Manual evaluation:
  - Manually assigned scored between 1-4 to every (query,rewrite) pair, by Yahoo! Team
  - Scores 1-2: relevant
  - Scores 3-4: irrelevant
  - $$precision(q,m) = \frac{relevant\ rewrites\ of\ q\ that\ m\ provides}{number\ of\ rewrites\ of\ q\ that\ m\ provides}$$
  - $$recall(q,m) = \frac{relevant\ rewrites\ of\ q\ that\ m\ provides}{number\ of\ relevant\ rewrites\ of\ q}$$

Experiments: Metrics

- Query Coverage:
  - Absolute number of queries for which there is at least one rewrite
- Rewriting Depth:
  - Number of rewrites for a given query

Experiments: Metrics

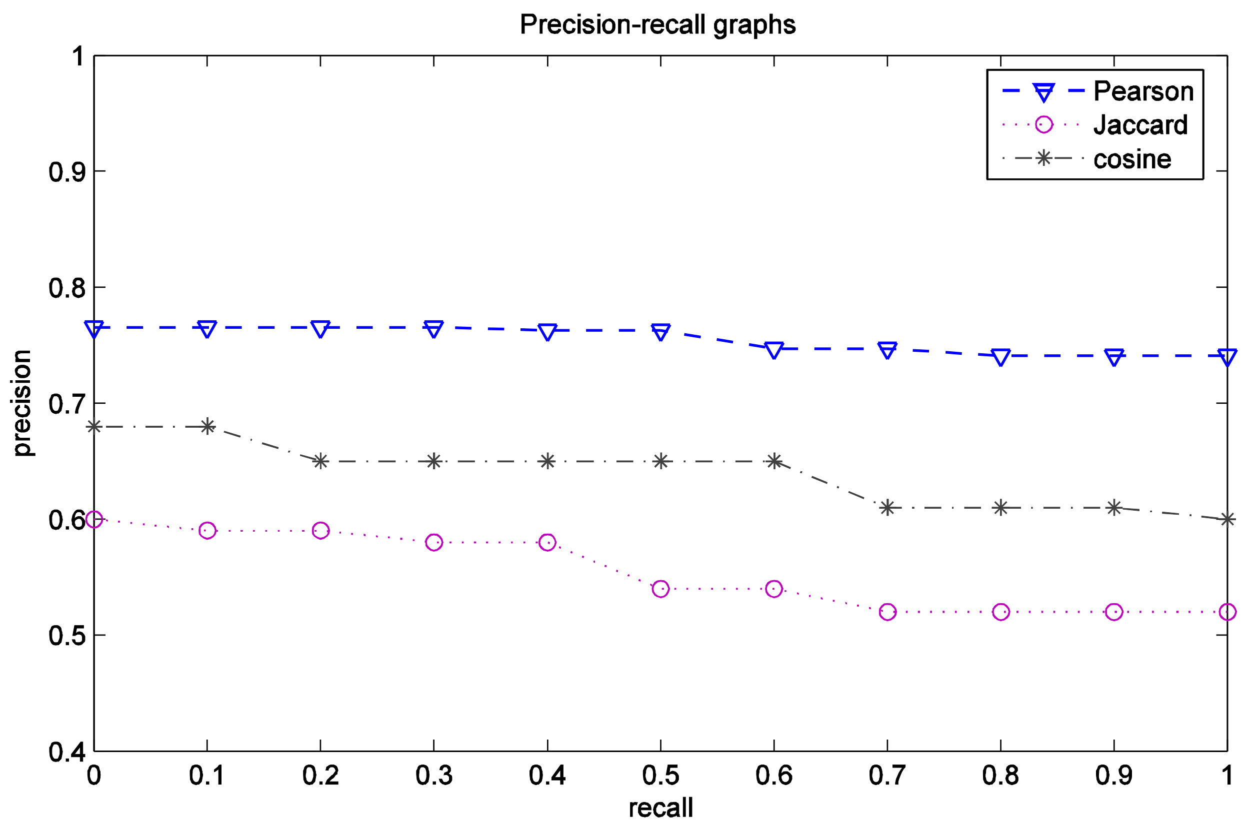
- Desirability:
  - Desirability function:

$$des(q_1,q_2) = \sum_{i \in E(q_1) \cap E(q_2)} \frac{1}{|E(q_2)|} \,.\, w(q_2,i)$$

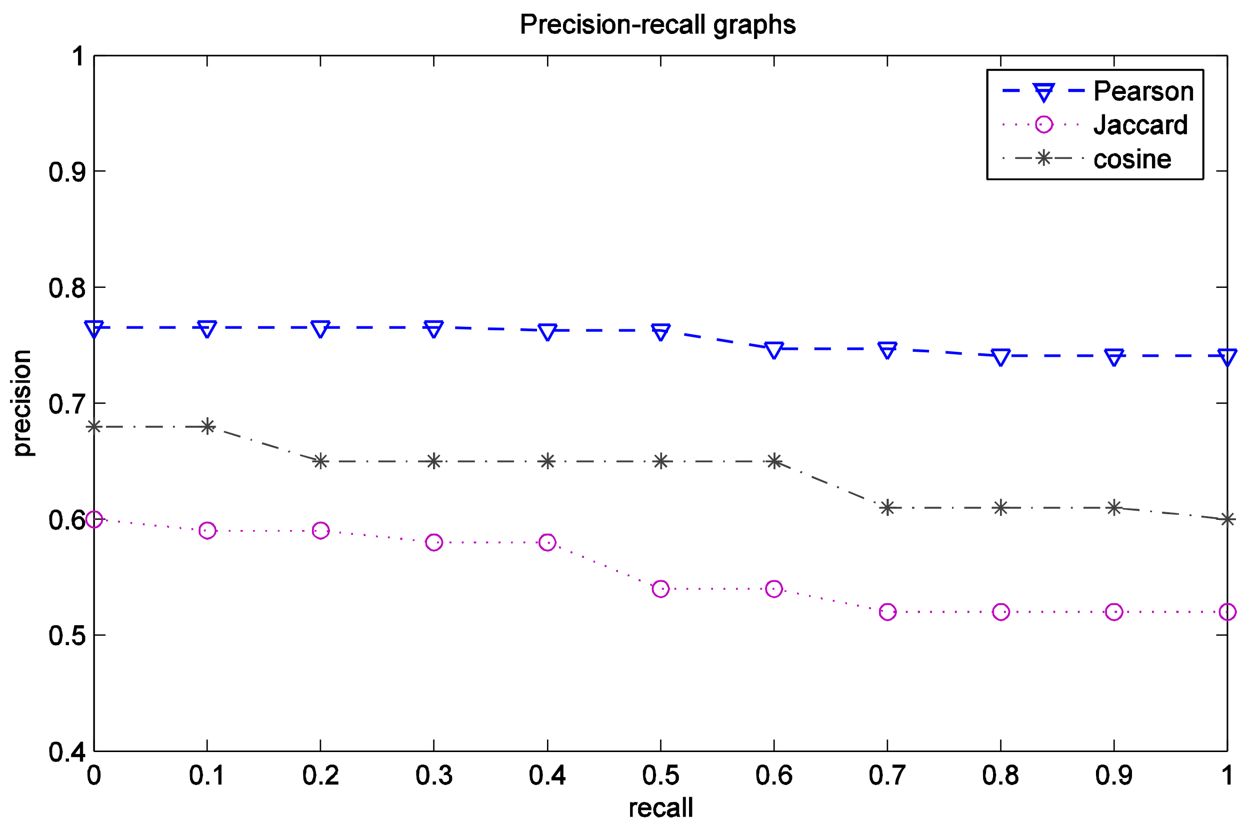
  - If $des(q_1,q_2) > des(q_1,q_3)$ then,

    $sim(q_1,q_2) > sim(q_1,q_3)$

Results: baselines

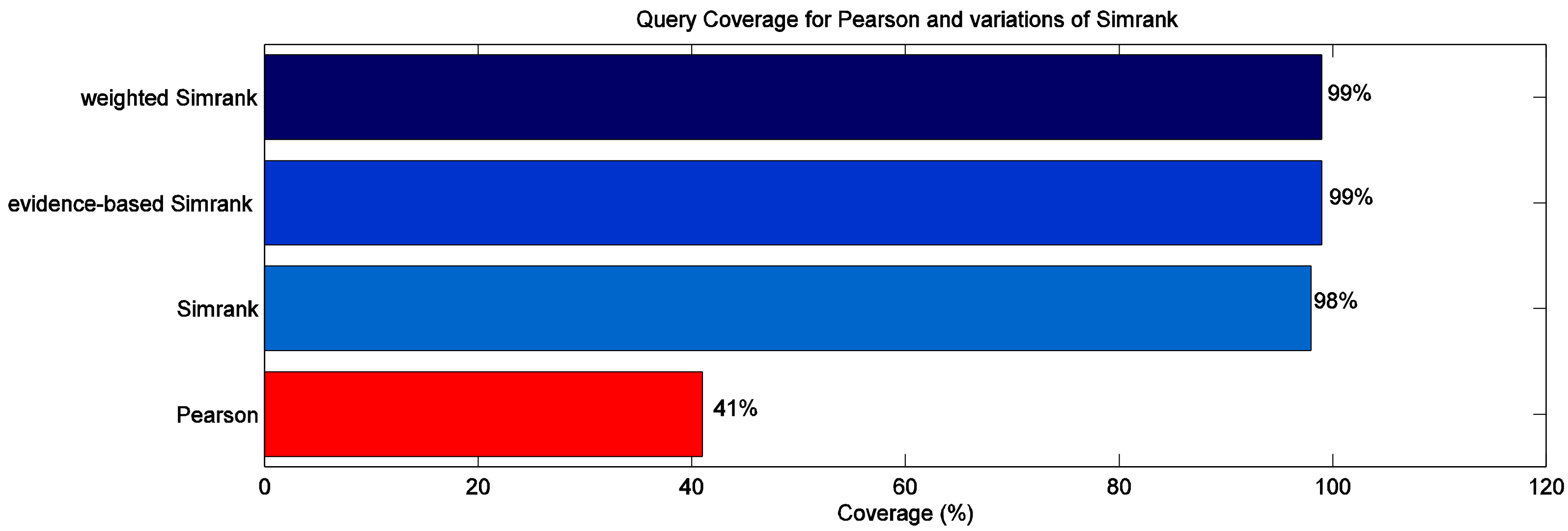


Results: baselines

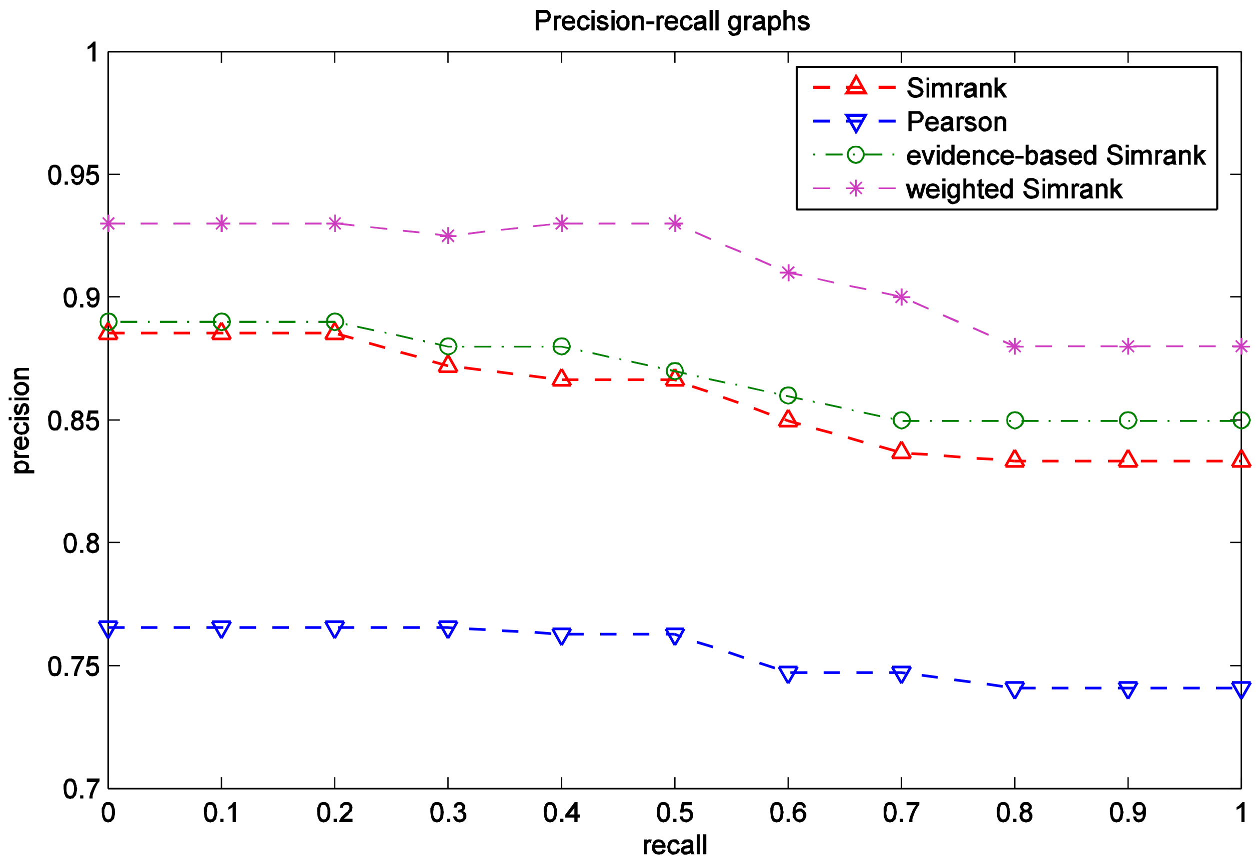


- Note: Pearson fares the best among baselines

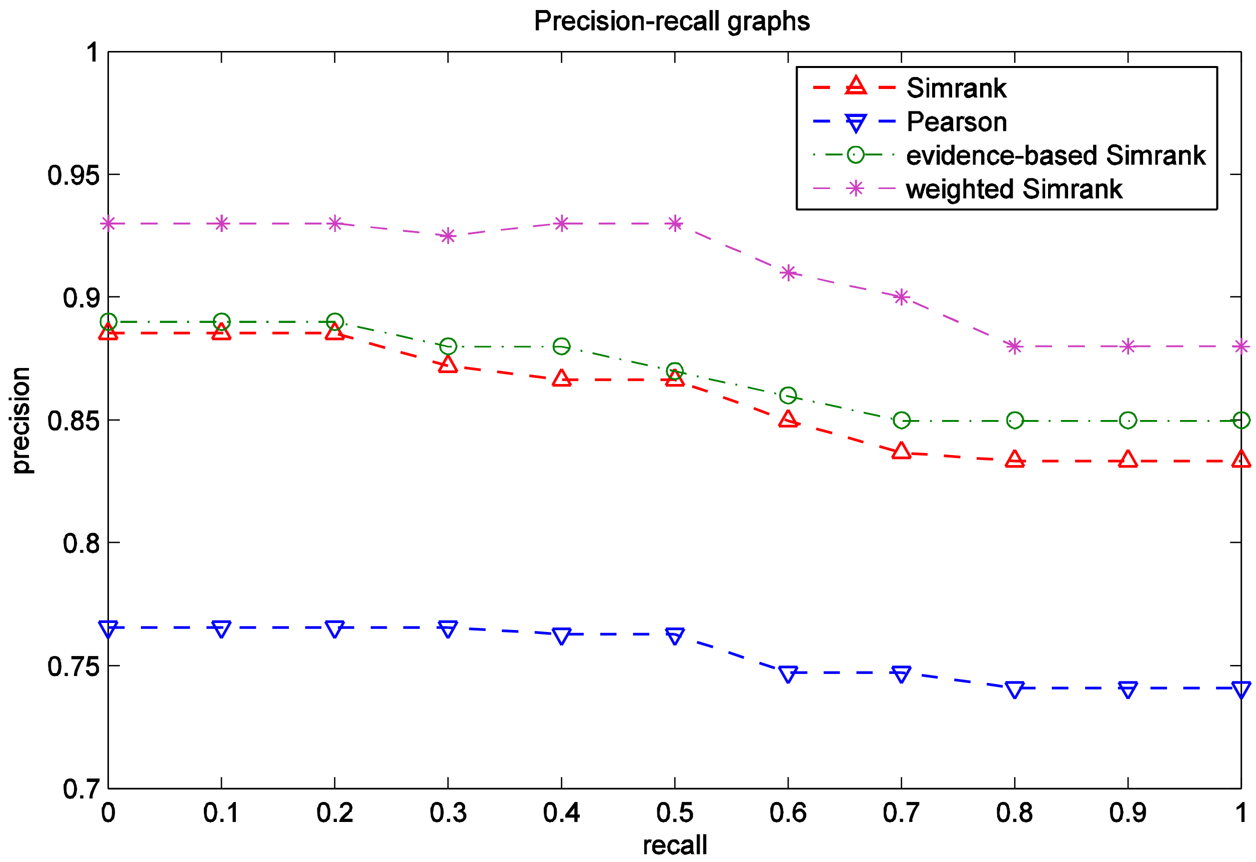
Results: query coverage



Results: precision-recall

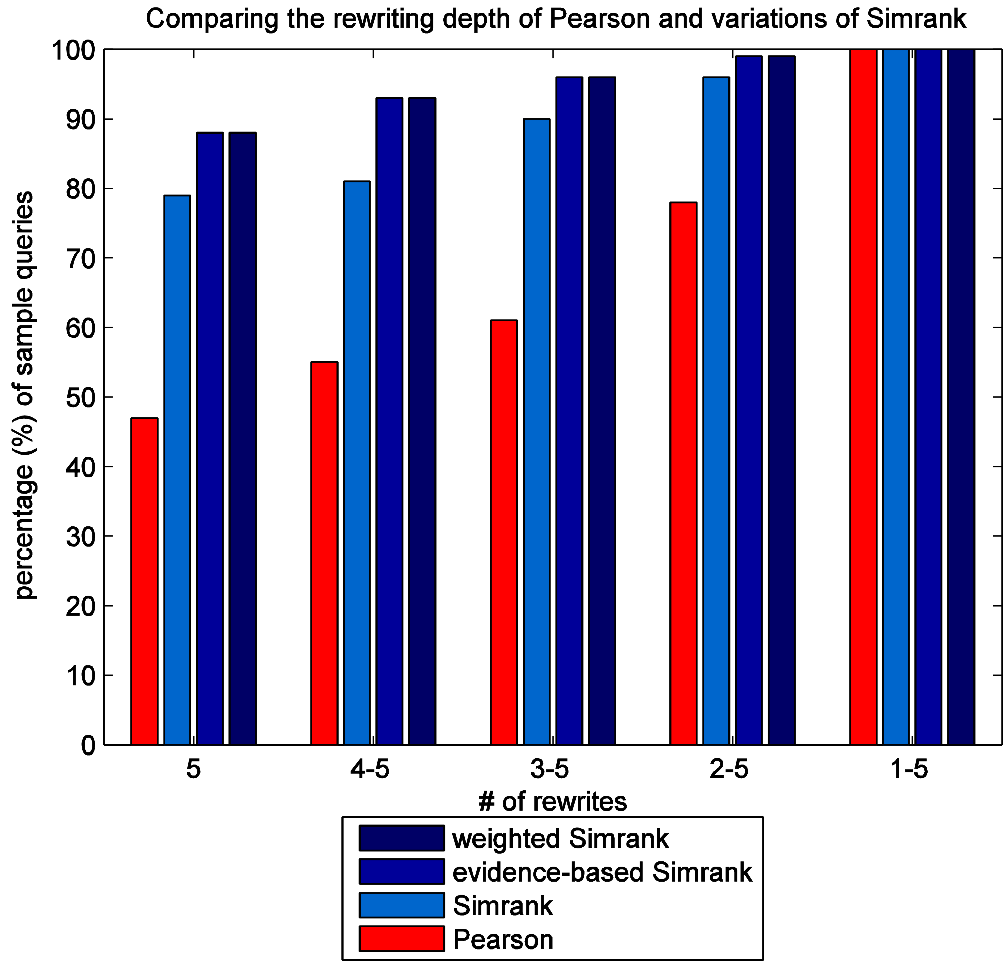


Results: precision-recall

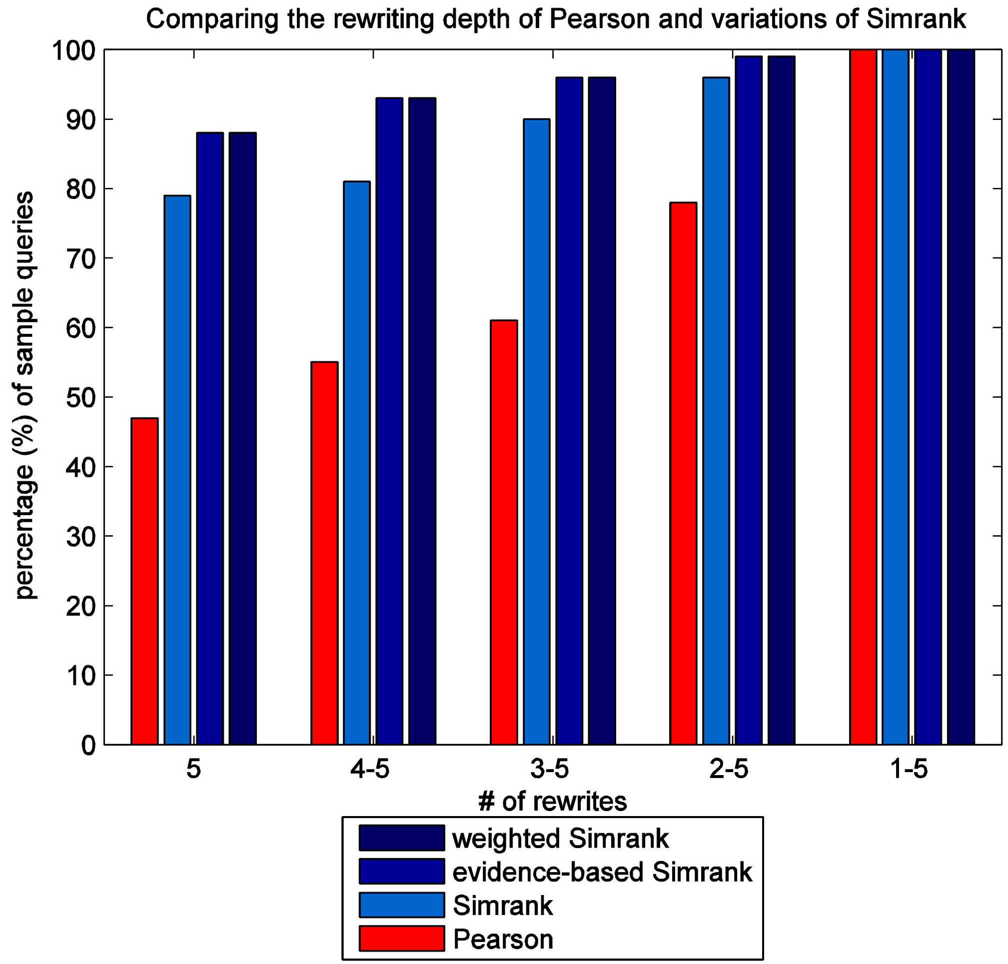


- Note: weighted simrank fares the best

Results: rewriting depth



Results: rewriting depth

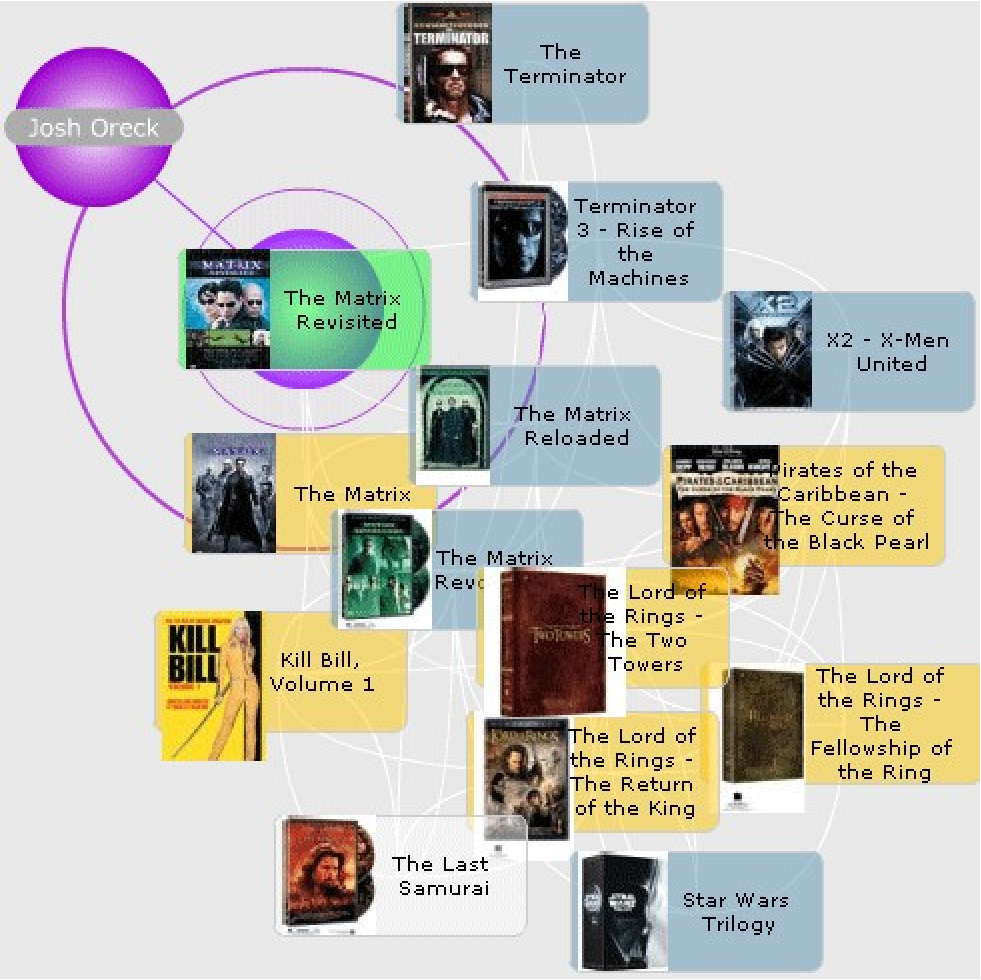


- Note: weighted simrank is among the best

Conclusion

- Two Simrank extensions
  - “evidence” supporting similarity
  - Weights of edges
- Weighted Simrank is overall the best
- Issues not addressed
  - Spam clicks
  - Semantic text-based similarities
  - Updating similarity scores with changes in click graph

Multi-partite?



What about more than two partitions?

THANKS