



Análise de Links

Danilo Carlos Gouveia de Lucena

Davi Hirafuji Neiva

Telmo de Menezes e Silva Filho



UNIVERSIDADE
FEDERAL
DE PERNAMBUCO



Roteiro

- Conceito
- Entender o porquê
- Métodos básicos
 - Page rank
 - HITS
 - Aplicações
- Classificação coletiva de dados conectados
- Análise de comunidades

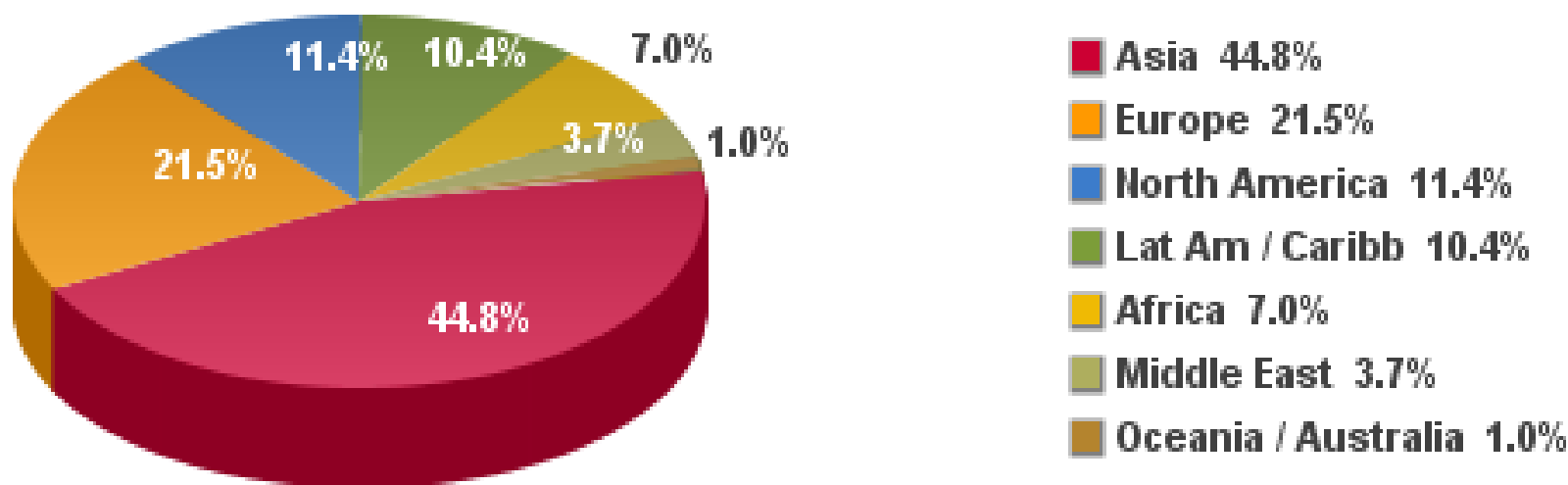


Internet

- Rede global de computadores
- Carrega uma massa de dados enorme
 - Documentos de hipertexto (WWW)
 - Redes P2P
 - Compartilhamento de arquivos (FTP)
 - Email, videos
 - etc ...



Internet Users in the World Distribution by World Regions - 2012 Q2



Source: Internet World Stats - www.internetworldstats.com/stats.htm

Basis: 2,405,518,376 Internet users on June 30, 2012

Copyright © 2012, Miniwatts Marketing Group



WWW

- World Wide Web
- Repositório universal de conhecimento
- Livre acesso
- Fonte distribuidora de informação
 - Blogs, tutoriais, vídeos, serviços, livros etc ...



Hiperlinks !!!





what is bitcoin - Google Search



https://www.google.com.br/search?q=google&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:pt-BR:official&client:



what is bitcoin

Web

Videos

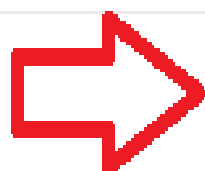
Images

News

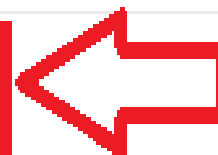
Maps

More ▾

Search tools



About 181,000,000 results (0.34 seconds)



Compre e Venda Bitcoin - MercadoBitcoin.com.br

Ad www.mercadobitcoin.com.br/ ▾

Negocie no Primeiro Site de Compra E Venda de Bitcoin do Brasil.

Ritcoin exchange - Markets.com

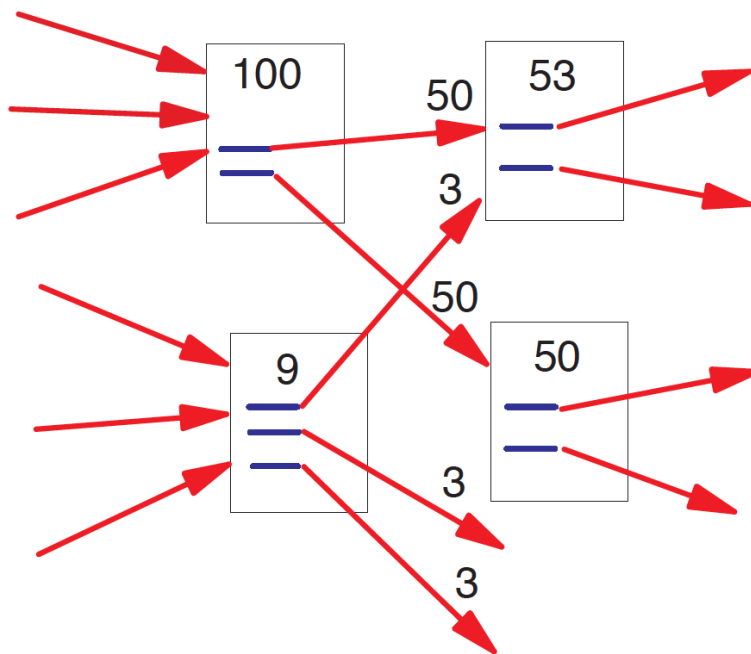
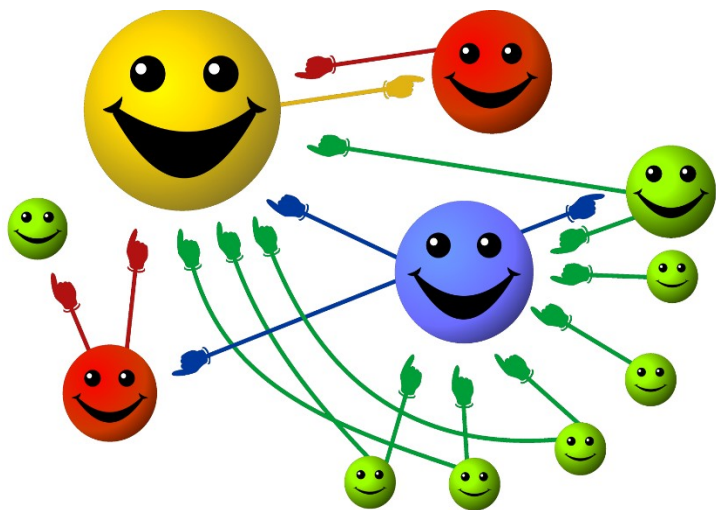


Page Rank

- Proposto por Sergey Brin e Lawrence Page 1998
- Algoritmo usado pela Google
- Rankear páginas de acordo com links que apontam para uma página ou clicks de uma página
 - Relevância das páginas
 - Importância das páginas



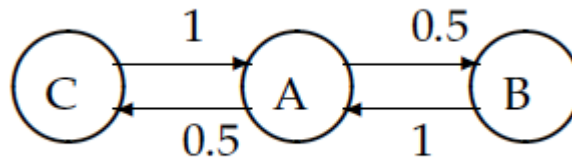
Page Rank





Page Rank cont.

- Teletransporte
- Markov chains
 - Processo que ocorre a cada teletransporte (algo randômico)



$$\begin{pmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$



- Se na linha não tiver nenhum 1, substitua cada elemento por $1/N$
- Divida cada 1 pela quantidade de 1's da linha
- Multiplique a matriz por $1-\alpha$
- Adicione $1/\alpha$ a cada elemento para obter a probabilidade



HITS - Hubs e Authorities

- Hubs = conjunto de informações que apontam para um especialista
- Authorities = o especialista
- Detectar especialistas através dos hubs

$$h(v) \leftarrow \sum_{v \mapsto y} a(y)$$

$$a(v) \leftarrow \sum_{y \mapsto v} h(y).$$



Aplicações

- <http://www.conversion.com.br/blog/open-site-explorer-ferramenta-de-analise-de-links/>



Classificação Coletiva de Dados Conectados

Collective Classification in Network Data

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor
University of Maryland, College Park

Brian Gallagher, Tina Eliassi-Rad
Lawrence Livermore National Laboratory



UNIVERSIDADE
FEDERAL
DE PERNAMBUCO



Classificação Coletiva

- Classificação coletiva é a classificação combinada de um conjunto de objetos usando os seguintes tipos de correlação:
 - Dados uma rede e um objeto o nessa rede:
 - Correlação entre o rótulo de o e seus atributos observados;
 - Correlação entre o rótulo de o e os atributos observados (incluindo rótulos) de objetos “vizinhos”;
 - Correlação entre o rótulo de o e os rótulos não observados de objetos vizinhos.



Classificação Coletiva

- Um problema clássico desse tipo de classificação é a atribuição, a páginas da internet, de classes que indicam seus tópicos.
- Assume-se que os rótulos de páginas interligadas sejam correlacionados.



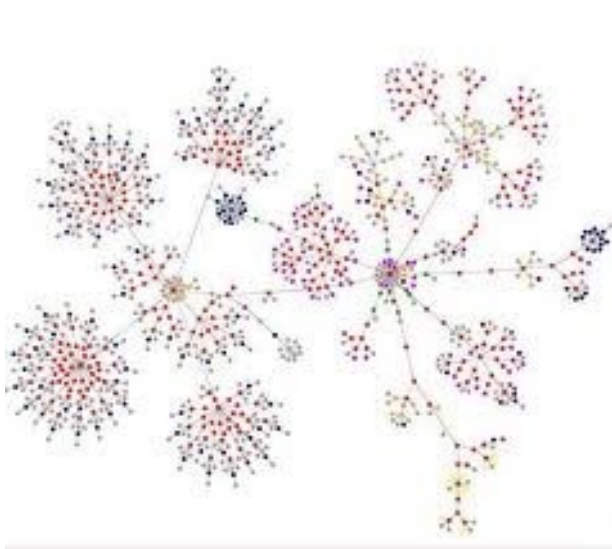
Classificação Coletiva

- Métodos de aprendizagem tradicionais ignorariam as correlações resultantes dessas interligações.
- Teoricamente certos métodos de inferência exata (por exemplo: eliminação de variáveis) poderiam ser empregados para Classificação coletiva.



Classificação Coletiva

- Problema: inferência exata é NP-hard.
 - Complexidade exponencial.
- A solução é buscar algoritmos de inferência aproximada.





Classificação Coletiva: Notação

- Num problema de Classificação coletiva, são dados:
 - Um conjunto de nós $V = \{V_1, \dots, V_n\}$
 - Uma função de vizinhança N , onde $N_i \subseteq V \setminus \{V_i\}$
 - V é dividido em dois conjuntos de nós:
 - X é o conjunto de nós cujos valores são conhecidos (variáveis observadas)
 - Y é o conjunto de nós cujos valores precisam ser determinados

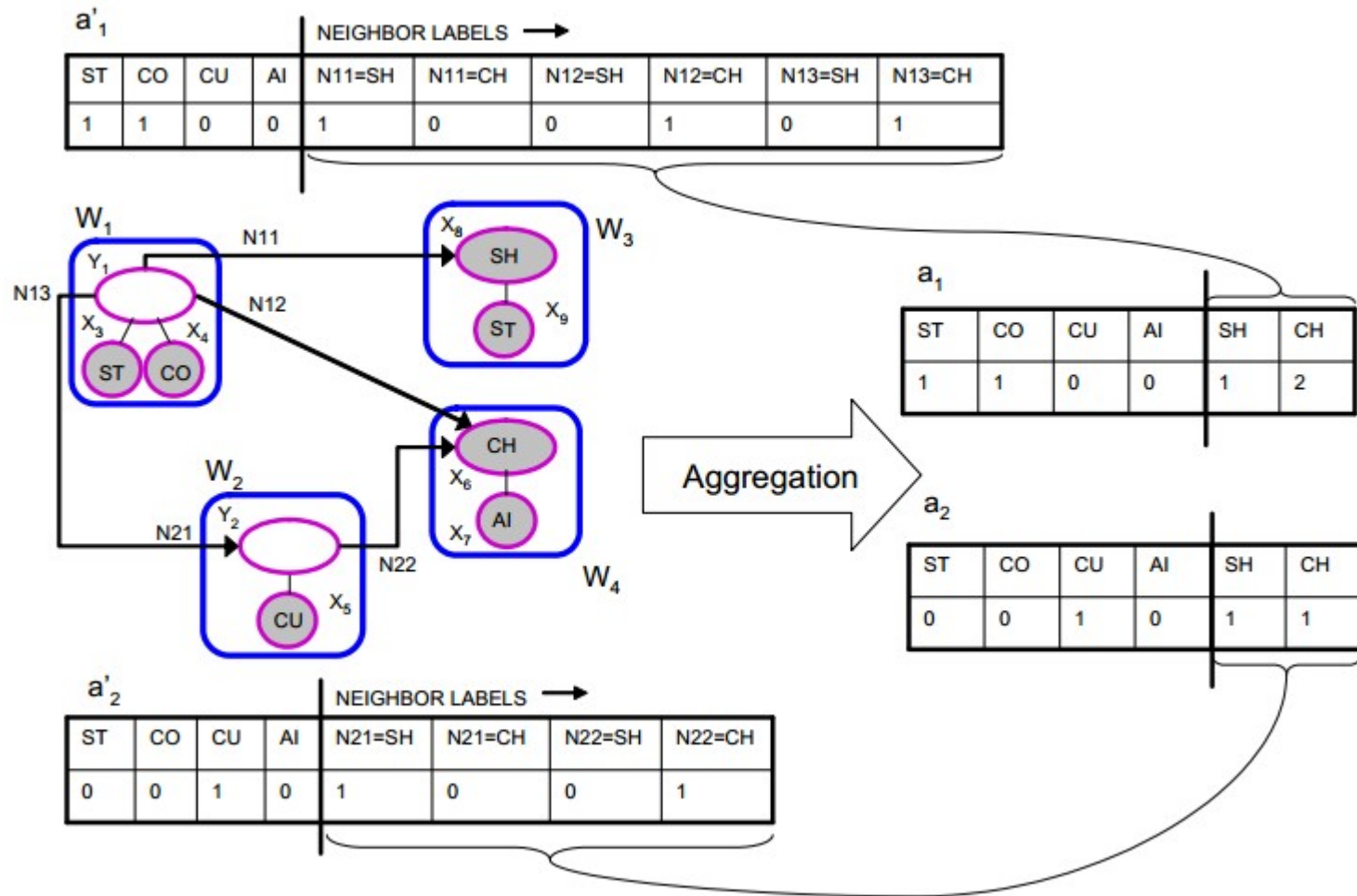


Classificação Coletiva: Notação

- Dados $X \subseteq V$, $Y \subseteq V$ e N , a tarefa do algoritmo escolhido será:
 - Classificar Y_i com um rótulo L_i retirado de um conjunto de rótulos $L = \{L_1, \dots, L_q\}$



Classificação Coletiva: Exemplo





Classificação Coletiva: Abordagens

- Existem duas abordagens principais para resolver o problema da Classificação coletiva:
 - Classificadores condicionais locais:
 - **Algoritmo de classificação iterativa (ICA)**
 - Amostragem de Gibbs (GS)
 - Otimização de função objetivo global:
 - **Propagação de crença em laço (LBP)**
 - Rotulação por afrouxamento de campo-médio (MF)



Classificação Iterativa

- Considere um nó $Y_i \in Y$ e suponha que sabemos os valores de todos os nós em sua vizinhança N_i .
- ICA assume que existe um classificador f que retorna o melhor valor de rótulo para Y_i , retirado do conjunto de rótulos L e baseado nos valores de N_i .



Classificação Iterativa

- ICA é flexível em relação ao classificador f .
- Infelizmente, na prática é raro que todos os valores de N_i sejam conhecidos.
- Por isso o processo precisa ser repetido iterativamente.

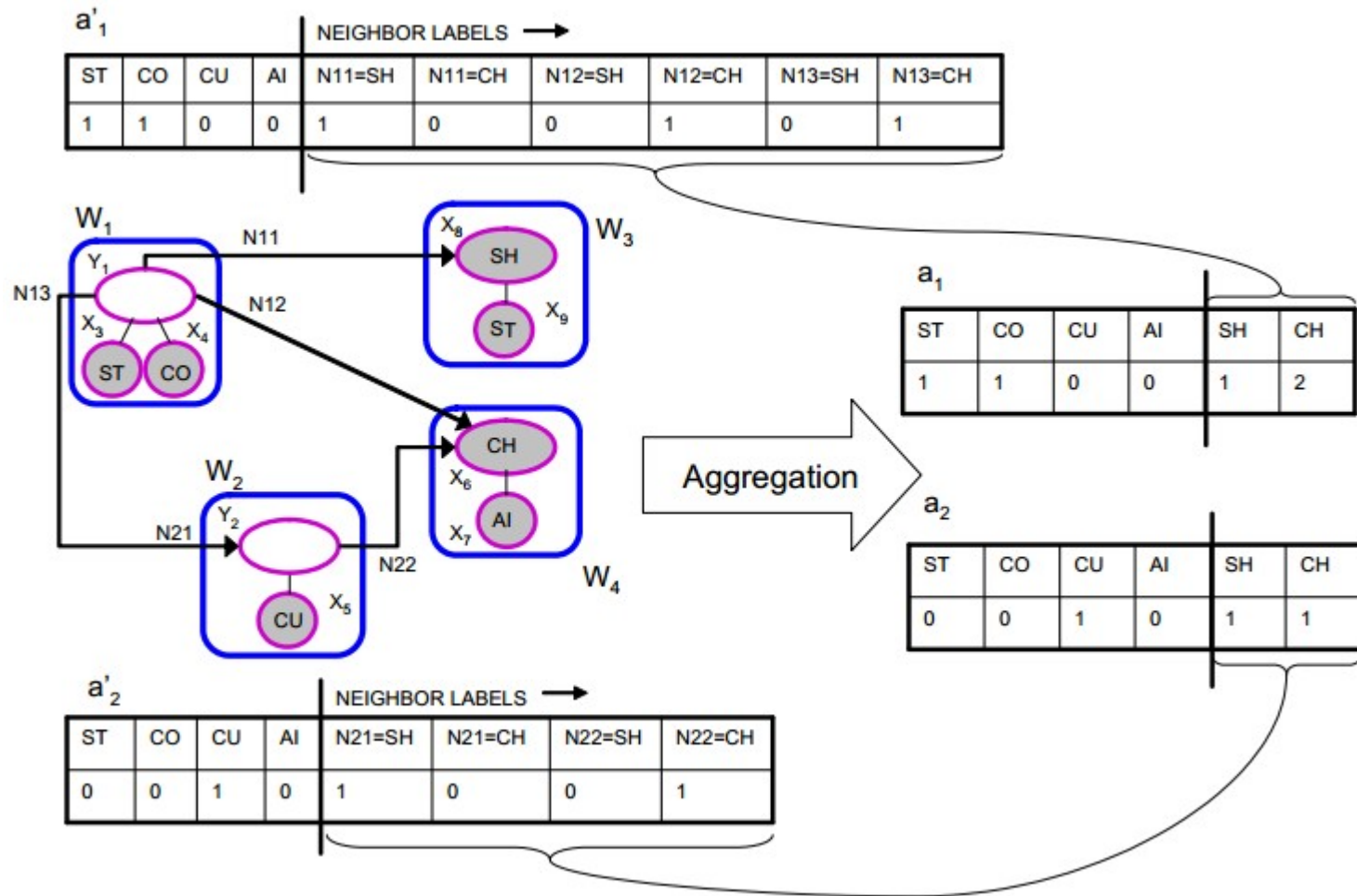


Classificação Iterativa

- A cada iteração, cada Y_i recebe um rótulo baseado nas melhores estimativas atuais para N_i e no classificador f .
- As iterações continuam até os rótulos estabilizarem.



Classificação Iterativa: Agregação





Classificação Iterativa: Fluxograma

Inicialização:
Compute os
rótulos dos
nós Y_i usando
apenas
 $X_i \in N_i$.

Iterações:
Compute os
rótulos dos
nós Y_i usando
os valores
atuais de N_i .

Estabilizou = False

Retorne os
rótulos
encontrados.

Estabilizou = True



Propagação de Crença em Laço

- O objetivo agora é otimizar uma função de objetivo global.
- O algoritmo trabalha sobre um Campo Aleatório de Markov (MRF) pareado.
 - Um MRF pareado é definido por um par $\langle G, \Psi \rangle$, onde G é um grafo e Ψ é o conjunto de potenciais de clique de G .



Propagação de Crença em Laço

- Propagação de Crença em Laço é um algoritmo que tem os seguintes passos:
 - Cada $Y_i \in Y$ comunica “mensagens” a seus vizinhos, até que essas mensagens estabilizem.
 - Quando as mensagens estabilizam, é possível calcular as probabilidades de que cada Y_i receba um rótulo $y_i \in L$.



Propagação de Crença em Laço

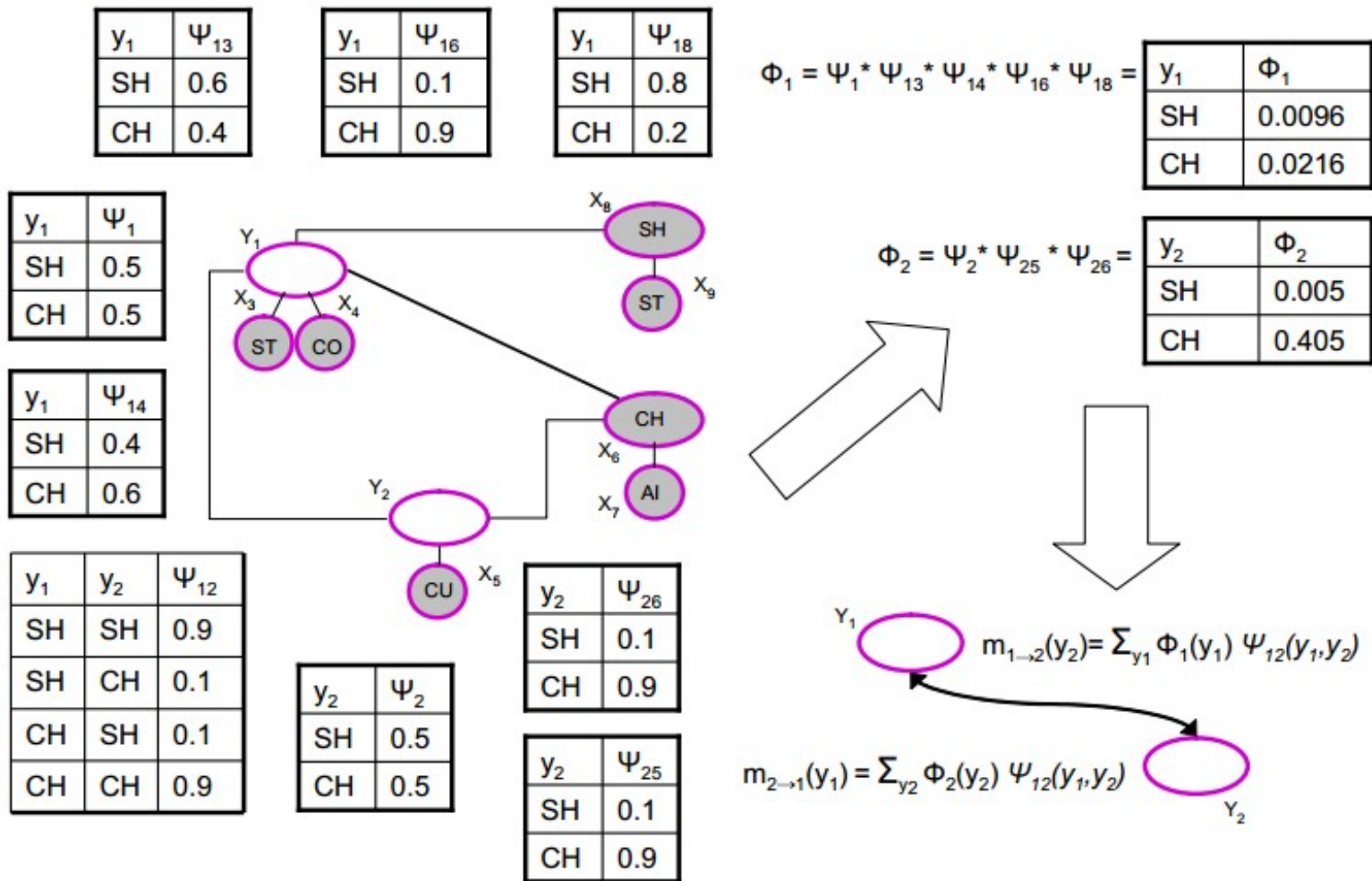
- As mensagens são baseadas nos potenciais de clique que envolvem cada Y_i e em mapeamentos (ψ) dos rótulos de L para valores reais não negativos.

$$m_{i \rightarrow j}(y_j) = \alpha \sum_{y_i \in \mathcal{L}} \psi_{ij}(y_i, y_j) \phi_i(y_i) \prod_{Y_k \in \mathcal{N}_i \cap \mathcal{Y} \setminus Y_j} m_{k \rightarrow i}(y_i), \quad \forall y_j \in \mathcal{L} \quad (1)$$

$$b_i(y_i) = \alpha \phi_i(y_i) \prod_{Y_j \in \mathcal{N}_i \cap \mathcal{Y}} m_{j \rightarrow i}(y_i), \quad \forall y_i \in \mathcal{L} \quad (2)$$



Propagação de Crença em Laço





Análise de Comunidades

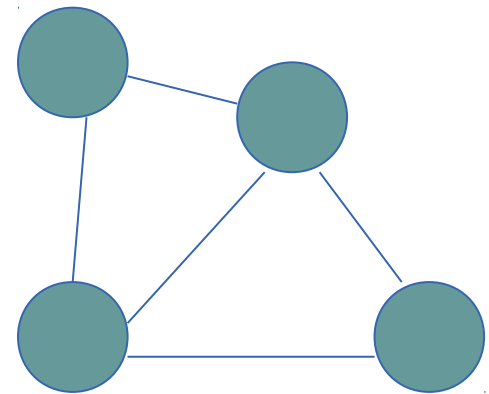
Community Structure in Large Networks: Natural Cluster Sizes
and the Absence of Large Well-Defined Clusters ^{*}

Jure Leskovec [†] Kevin J. Lang [‡] Anirban Dasgupta [†] Michael W. Mahoney [§]



Análise de Comunidades

- Definir e identificar clusters ou comunidades
- Redes sociais e/ou de informação





Proposta

- Nós representam entidades sociais
- Ligações representam interações

Como identificar comunidades em redes sociais e informacionais que possuem uma grande quantidade de nós e interações?



Proposta

- Abordagens tradicionais: extrair conjuntos e interpretar comparando com o cenário real
- Melhor: aplicar algoritmos aproximativos para o problema de particionamento de grafo
- Em função de estatísticas de tamanho e de propriedades da estrutura

{ Network Community Profile Plot (NCP)



NCP

- Melhor comunidade: medida de condutância
- Algoritmo iterativo
- Testes feitos em 100 redes reais
- Nova visão da área
- Não existe uma definição de comunidade

{ Conjunto de nós com melhores conexões entre si do que entre os demais nós da rede



Como definir comunidades?

- (1) Modelar os dados como um grafo com nós e interações
- (2) Existem conjuntos de nós que interagem melhor entre si do que com o resto da rede
- (3) Métrica para formalizar a ideia de grupos
- (4) Escolher algoritmo para encontrar os nós
- (5) Avaliar (comparar com o mundo real; estatísticas ...)



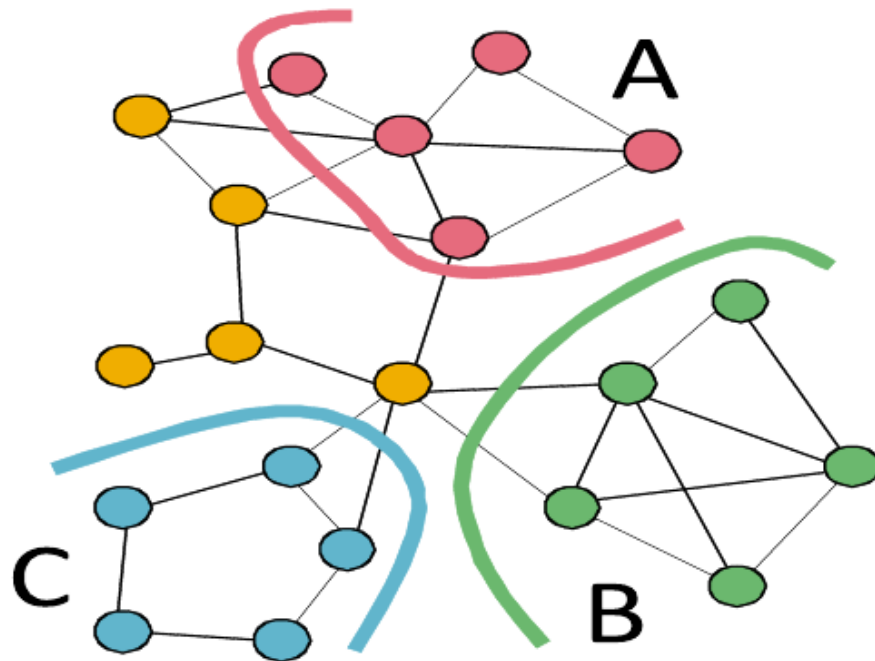
Como definir comunidades?

- Uma boa comunidade deve ter um baixo valor de condutância
- Mais nós internos e poucos nós apontando para o resto da rede
- No ponto (5) ao invés de comparar com um cenário real, o NCP analisa as propriedades das redes
- NP-Hard! Mas usa algoritmos aproximativos

{ Resolver o Minimum Conductance Cut Problem



Como o NCP funciona?



Determina o processo de detecção de comunidades e as propriedades pela análise das várias redes



Resultados?

> 100 nós

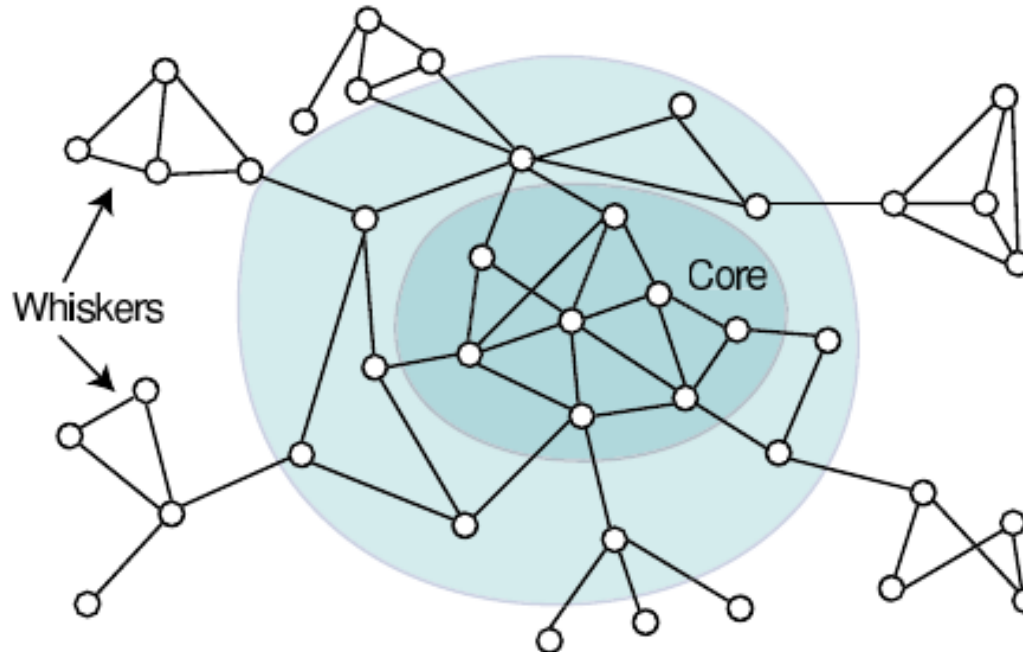
- NCP aumenta
- Observa-se que as comunidades vão se “misturando” com o resto da rede e gradualmente desaparecem quando ficam muito grandes

< 100 nós

- Boas comunidades (baixa condutância). NCP é ideal
- Comunidades menores podem ser combinadas em comunidades maiores



Tamanho x NCP



- Whiskers são nós conectados ao resto do grafo por uma ligação simples.
- São responsáveis pelo aumento do NCP. Diminuem a chance de detectar boas comunidades.



Redes Analisadas

Delicious Nós: 147,567 Links: 301,921	Web-Google Nós: 855,802 Links: 4,291,352	Actor Movies IMDB Nós: 2,076,978 Links: 5,847,693
Flickr Nós: 404,733 Links: 2,110,078	eDonkey Nós: 5,792,297 Links: 147,829,887	Amazon0302 Nós: 262,111 Links: 899,792
Actor IMDB Nós: 883,963 Links: 27,473,042	Netflix Nós: 497,959 Links: 100,480,507	IpTraffic Nós: 2,250,498 Links: 21,643,497



Definição: Condutância

Definicion 1 *Given a graph G with adjacency matrix A the conductance of a set of nodes S is defined as:*

$$\phi(S) = \frac{\sum_{i \in S, j \notin S} A_{ij}}{\min\{A(S), A(\overline{S})\}}, \quad (6)$$

where $A(S) = \sum_{i \in S} \sum_{j \in V} A_{ij}$, or equivalently $A(S) = \sum_{i \in S} d(i)$, where $d(i)$ is a degree of node i in G . Moreover, in this case, the conductance of the graph G is:

$$\phi_G = \min_{S \subset V} \phi(S). \quad (7)$$

- O grafo $G = (V, E)$
- Definir a condutância
- S é um conjunto de V
- O complemento de S



Definição: NCP

Definition 2 Given a graph G with adjacency matrix A , the network community profile plot (NCP plot) plots $\Phi(k)$ as a function of k , where

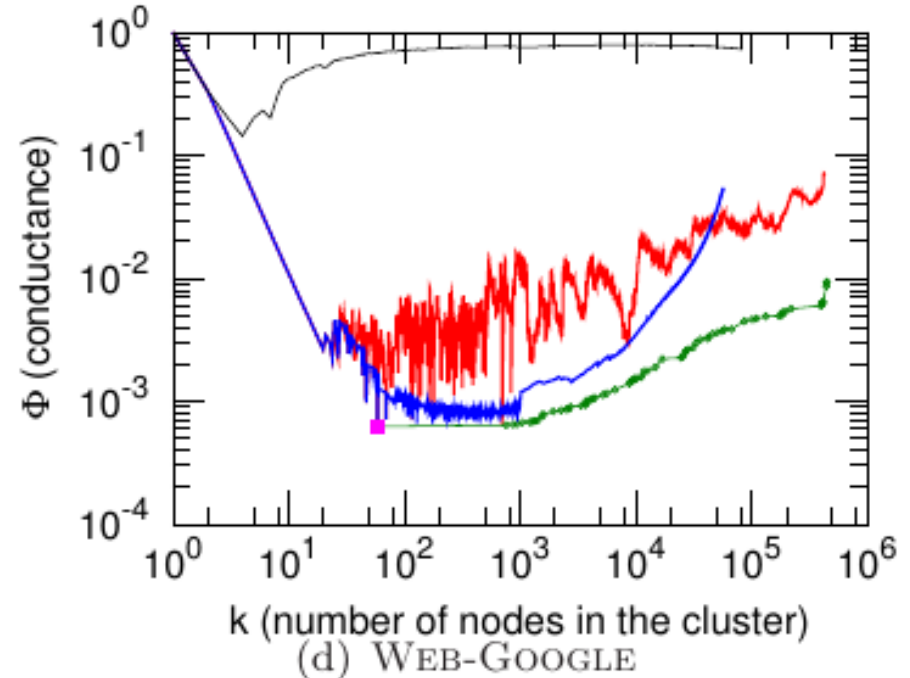
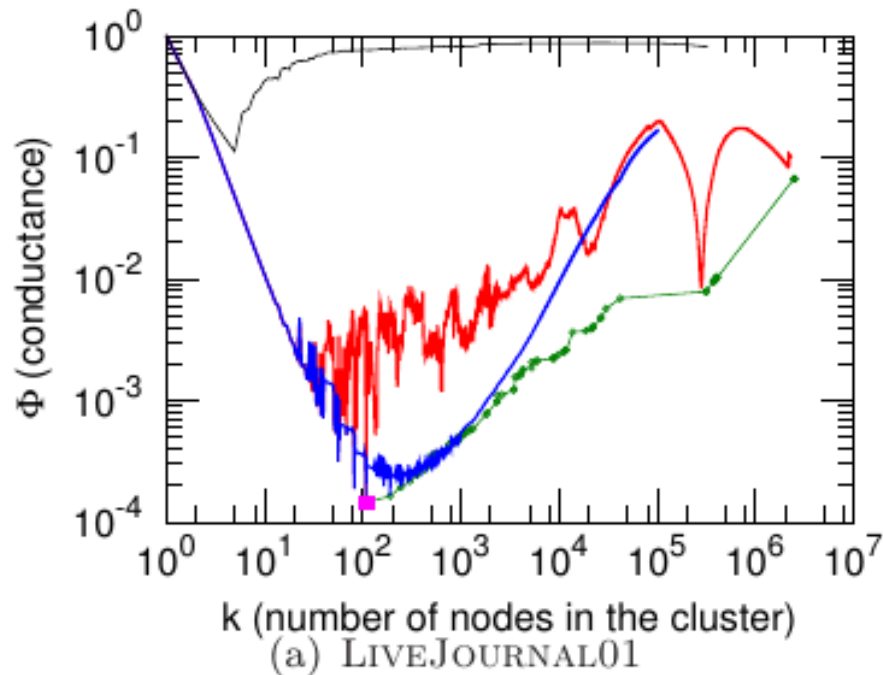
$$\Phi(k) = \min_{S \subset V, |S|=k} \phi(S), \quad (8)$$

where $|S|$ denotes the cardinality of the set S , and where the conductance $\phi(S)$ of S is given by equation (6).

- O resultado da função NCP é a medida da qualidade da melhor comunidade possível para uma rede grande
- K é o valor de cardinalidade (ou tamanho) do conjunto
- A quantidade possível de tamanhos torna o problema intratável. Aplica-se então algoritmos aproximativos

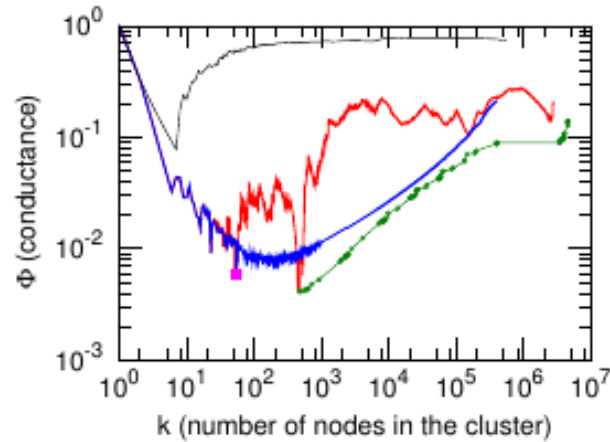


Aplicando os algoritmos

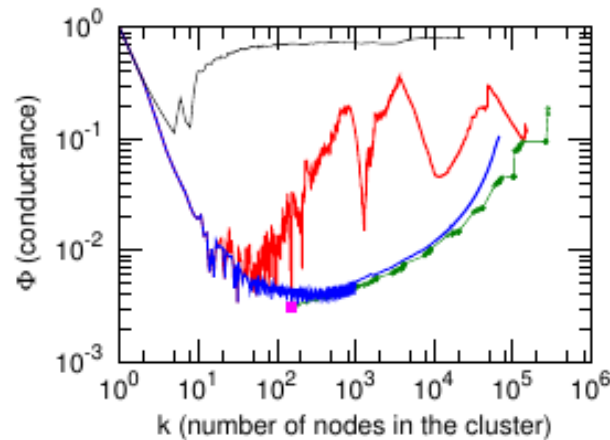




Aplicando os algoritmos



LINKEDIN



FLICKR

Local Spectral —
Metis+MQI —○—

Rewired network —
Bag of whiskers —



Vale a pena?

- Para ter uma visão detalhada da área de criação de comunidades em redes grandes
- Quando é possível transformar dados em grafos (BigData, repositórios NoSQL, indexadores web)
- A abordagem considera apenas aspectos da estrutura, não valida com o “mundo real”

{ O estudo da matemática e teoria dos grafos relacionada é bem interessante

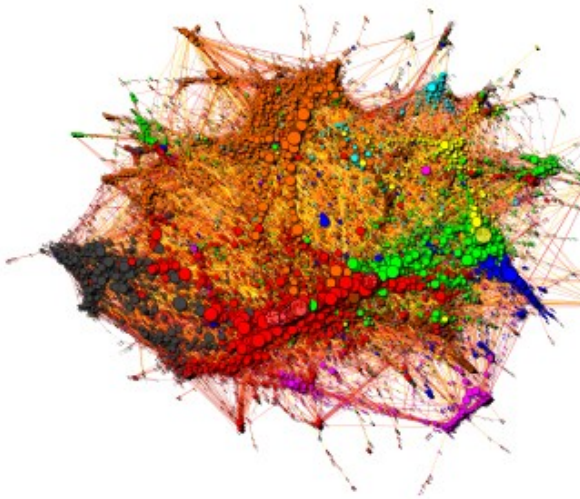


Usos?



- Comunidades
- Jogos
- Ads

My Music Habits



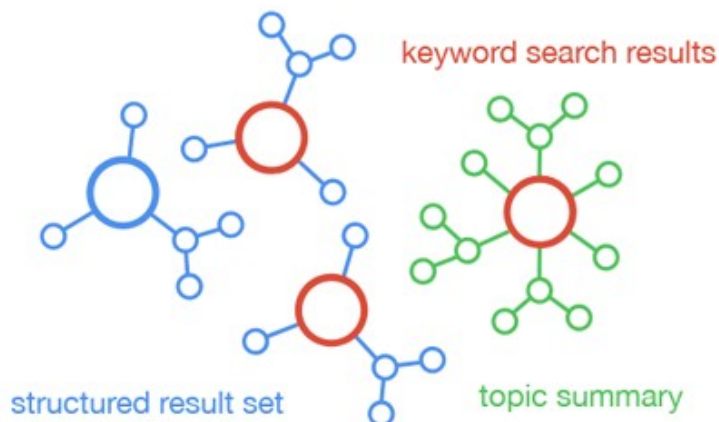
- Last.fm App
- Detectar grupos
- Sistema de recomendação?



Usos?



- Sistemas de recomendação
- Análise de visualizações
- Perfil



- Freebase.com
- Base de dados em grafos
- Apps semânticas



Freebase



Tap into the power of Google's Knowledge Graph with Freebase data

Freebase is a free, knowledge graph with millions of people, places, and things.

Francis Ford	
Select an item from the list:	
Francis Ford Coppola	Film director
Francis Ford	Film director
Jeremy Clarkson	Film director
Phillip Ford	Film director

```
ns:m.012rkqx ns:type.object.type
ns:m.012rkqx ns:type.object.name
ns:m.012rkqx ns:type.object.type
ns:m.012rkqx ns:type.object.key
ns:m.012rkqx ns:type.object.type
ns:m.012rkqx key:authority.musicbr
ns:m.012rkqx key:authority.musicbr
```

	Abraham Lincoln /m/0gzh Date of birth: 1809-02-12 Date of death: 1865-04-15 Place of birth: Hodgenville President number: 16 Vice president: Andrew Johnson,
---	---

Identify Freebase topics in your app

Use the [Search Widget](#) to easily lookup and disambiguate entities in your application.

Download the complete graph as RDF

Every fact and entity is available as an [RDF Dump](#) which makes it easy to analyze all of Freebase on your own computer.

Get images, descriptions and note-worthy facts

The [Topic API](#) allows you to get all the facts for an entity in JSON or RDF.



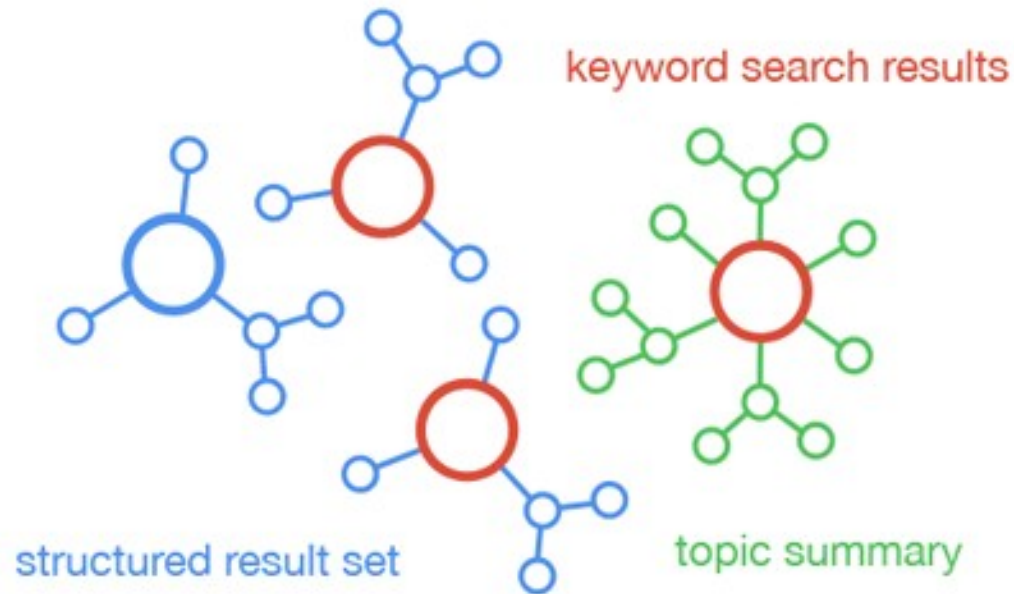
Freebase

- 2007: Início do projeto
- 2010: Comprado pelo Google
- Base de dados colaborativa
- Estruturado: tópicos, tipos, propriedades
- Tópicos possuem ligações com outros tópicos
- Anotados com propriedades

+ 2.5 bilhões de fatos
+ 43 milhões de tópicos



Freebase



- Grafos!
 - Metaweb Query Language
 - RDF (Resource Description Framework) – SPARQL?



Freebase



Topic

Bob Dylan ^{en}

mid: /m/01vrncs notable type: /celebrities/celebrity on the web:  wikipedia.org ▼

Bob Dylan is an American musician, singer-songwriter, artist, and writer. He has been an i dates from the 1960s, when he was an informal chronicler and a seemingly reluctant figure. "Are a-Changin'", became anthems for the US civil rights and anti-war movements. Leaving radically altered the parameters of popular music in 1965. His recordings employing elect incorporated a variety of political, social, philosophical, and literary influences. They defied by the performance style of Little Richard, and the songwriting of Woody Guthrie, Robert J

- Tópicos: entidades físicas, filmes, músicas, etc.
- Tipos: músico, escritor, pintor, escultor
- Propriedades: todas as músicas compostas pelo Bob Dylan



Freebase

Query

```
{
  [{
    "type": "/film/film",
    "only:starring": {
      "actor": {
        "name": "Kevin Spacey"
      }
    },
    "name": null
  }]
}
```



Topic

Kevin Spacey^{en}

mid: /m/048lv notable type: /business/board_member

Kevin Spacey, CBE is an American actor, director, screenwriter, and film producer. He gained critical acclaim in the early 1990s for his roles in the films L.A. Confidential and Seven. Since 2003, he has been artistic director of the Old Vic theatre.

Frank Underwood?

Resultado JSON

```
{
  "result": [
    {
      "name": "American Beauty",
      "type": "/film/film",
      "only:starring": {
        "actor": {
          "name": "Kevin Spacey"
        }
      }
    },
    {
      "name": "Working Girl",
      "type": "/film/film",
      "only:starring": {
        "actor": {
          "name": "Kevin Spacey"
        }
      }
    }
  ],
}
```



Freebase

Freebase Triples

This dataset contains every fact currently in Freebase.

Total triples: 1.9 billion

22 GB gzip

Updated: Weekly

250 GB uncompressed

Data Format: [N-Triples RDF](#)

License: [CC-BY](#)

Freebase/Wikidata Mappings

The data has been created based on the Wikidata-Dump of October 28, 2013, and contains only those links that have at least two common Wikipedia-Links and not a single disagreeing Wikipedia-Link. Furthermore, the lines are sorted by the number of common Wikipedia-Links (although in Turtle this does not really matter).

Total triples: 2.1M

21.2 MB gzip

Updated: October 28, 2013

242.9 MB uncompressed

Data Format: [N-Triples RDF](#)

License: [CC0](#)

- Dados indexados do Google disponíveis
- Download gratuito
- APIs / Bibliotecas



Ideias?

Analizador semântico

- Coletar dados de bases de dados, processar o texto, separar possíveis tópicos
- Fazer buscas na base do Freebase
- Buscar ligações no grafo entre os tópicos
- Determinar validade semântica/relacionamento do conteúdo com o “mundo real”



Ideias?

“Hurricane, de Bob Dylan, é um protesto contra a prisão de Rubin Carter”

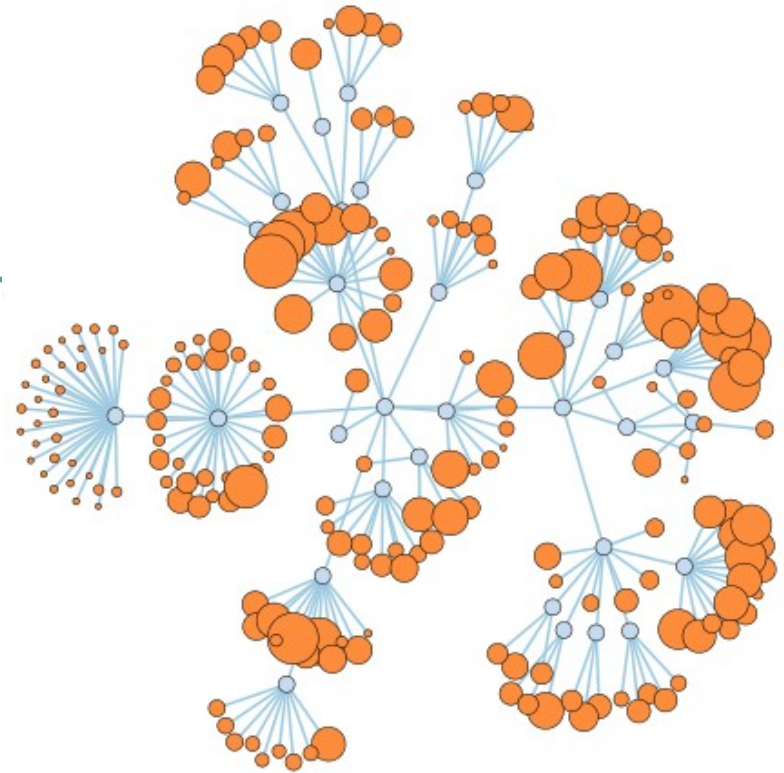
- Bob Dylan: Cantor, compositor
- Hurricane: Música do Bob Dylan
- Rubin Carter: Lutador de boxe. Preso injustamente
- Para cada tópico é criado um grafo
- Usar algoritmos para fazer buscas e identificar pontos de ligação entre os grafos
- Seria criado um “verificador semântico” automatizado
- Google indexa tudo?



Ideias?

Processar grafo em tempo r

- JSON
- D3.js
- Visualizar on-line
- Aplicar algoritmos





Sites

- freebase.com
- developers.google.com/freebase
- d3js.org