Max Planck Institute
for Human Development

From relational databases to linked data:R for the semantic web

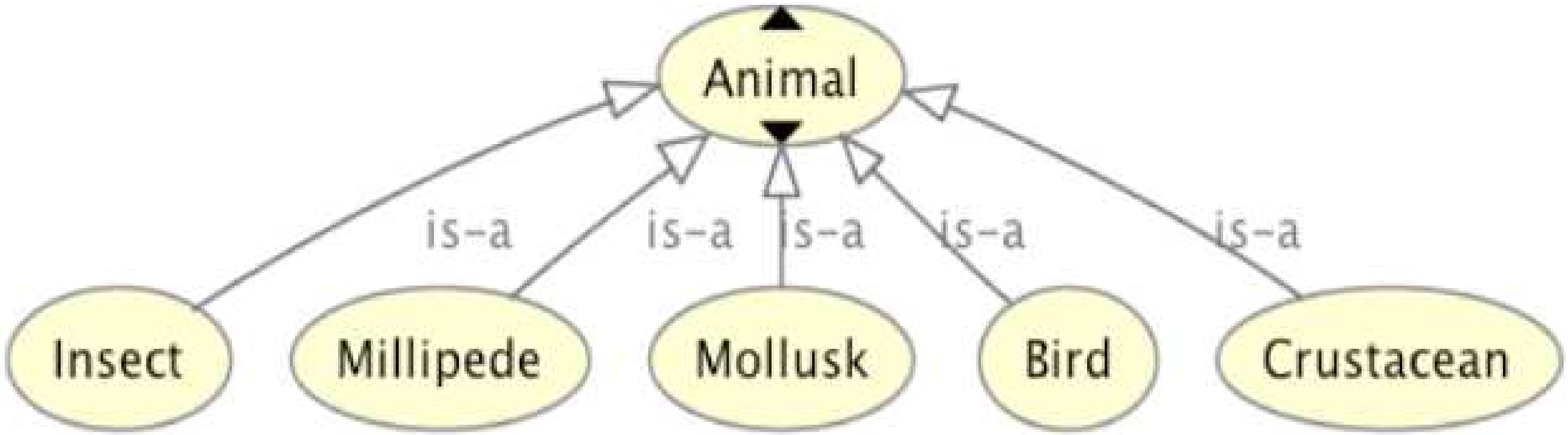
Jose Quesada,

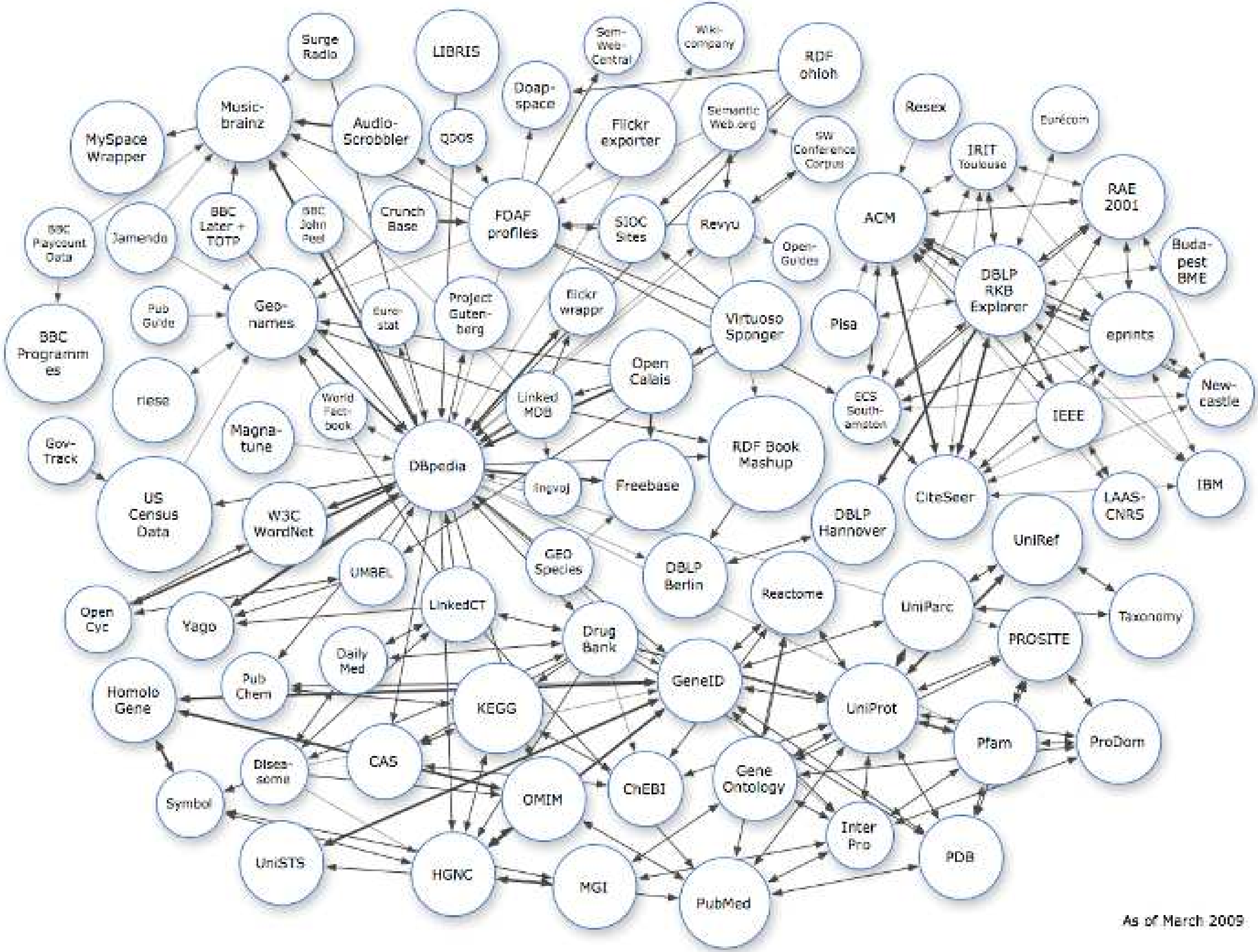
Max Planck Institute, Berlin

Who this talk targets

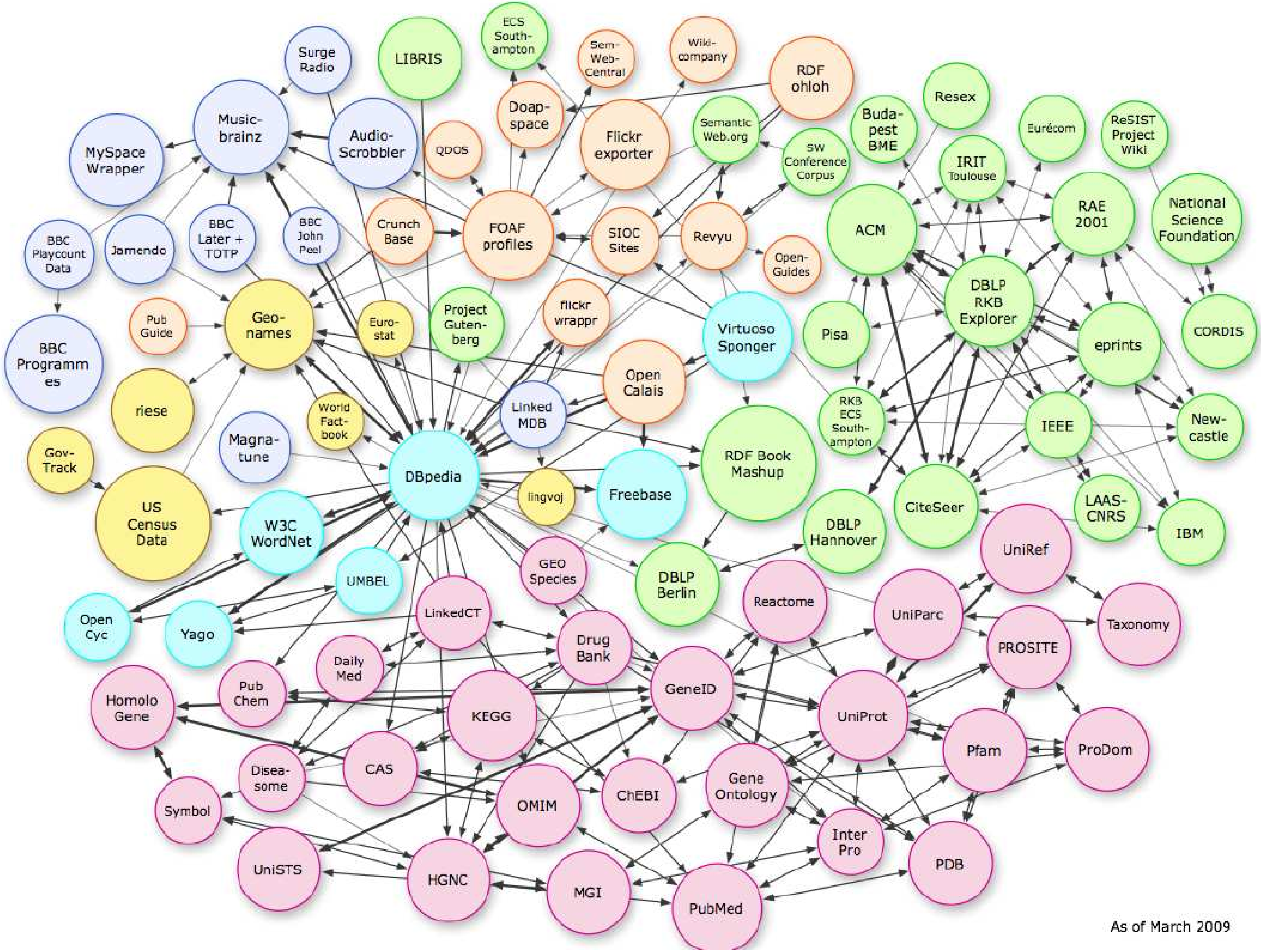
- You have big data; you use a database
- You have an evolving schema definition. Sometimes at runtime
- You are interested in alternative ways to present your data
- You would thrive by using data out there, if only they were more accessible

Semantic web





As of March 2009


Surge Radio
LIBRIS
ECS Southampton
Sem-Web-Central
Wiki-company
RDF ohloh
Music-brainz
MySpace Wrapper
Audio-Scrobbler
QDOS
Doap-space
Flickr exporter
Semantic Web.org
SW Conference Corpus
Resex
Buda-pest BME
Eurécom
ReSIST Project Wiki
IRIT Toulouse
National Science Foundation
RAE 2001
ACM
BBC Playcount Data
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
Open-Guides
DBLP RKB Explorer
eprints
CORDIS
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
Pisa
BBC Programmes
riese
World Fact-book
Linked MDB
Open Calais
RKB ECS South-ampton
IEEE
New-castle
Gov-Track
Magna-tune
DBpedia
lingvoj
Freebase
RDF Book Mashup
US Census Data
W3C WordNet
CiteSeer
LAAS-CNRS
IBM
DBLP Hannover
UMBEL
GEO Species
DBLP Berlin
UniRef
Open Cyc
Yago
LinkedCT
Drug Bank
Reactome
UniParc
Taxonomy
PROSITE
Daily Med
Homolo Gene
Pub Chem
KEGG
GeneID
UniProt
Pfam
ProDom
Disea-some
CAS
OMIM
ChEBI
Gene Ontology
Symbol
Inter Pro
PDB
UniSTS
HGNC
MGI
PubMed
As of March 2009

Credit: Jim Hendler

THE TWO TOWERS

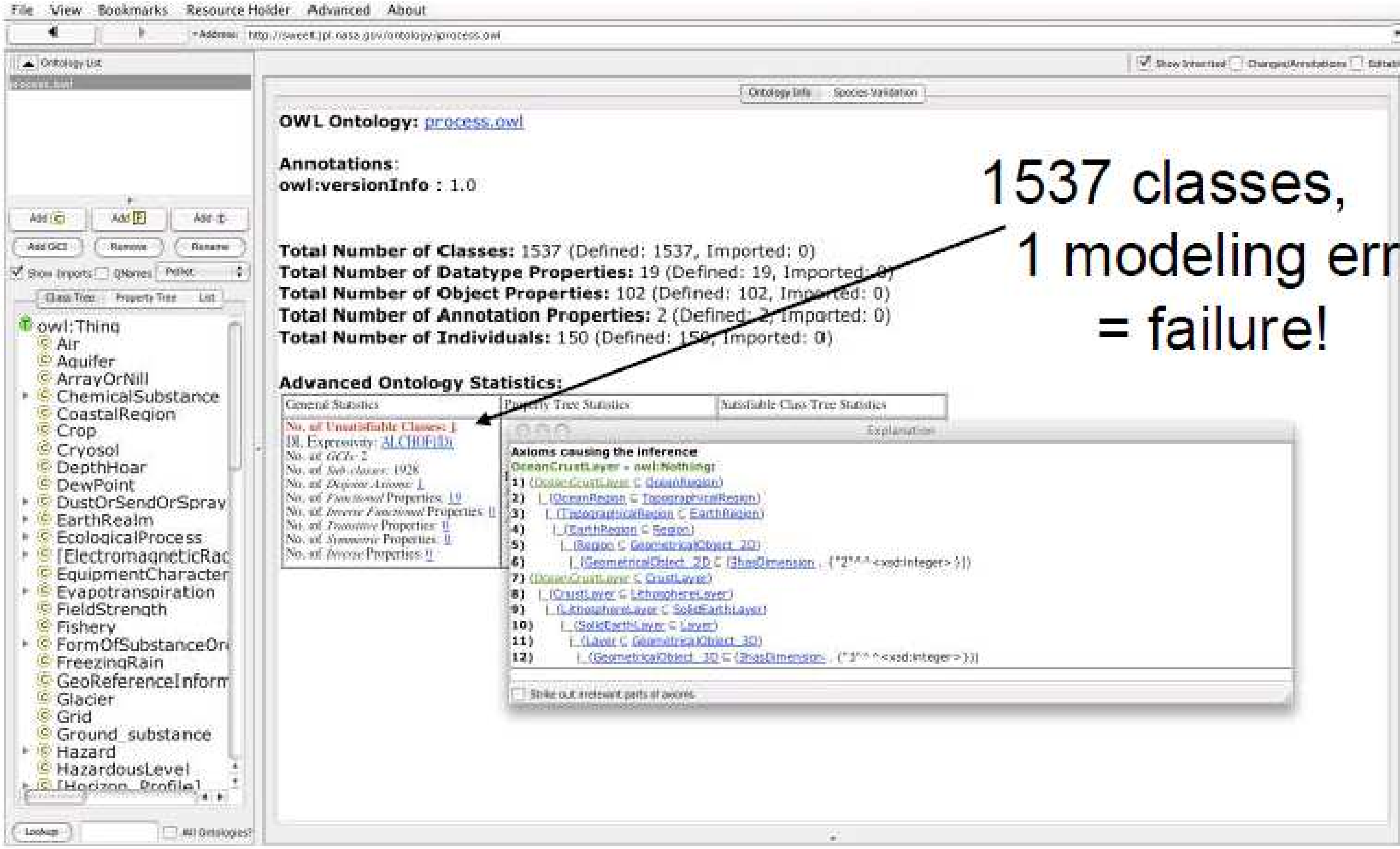
The *S*emantic web

- Ontology as Barad-dur (Sauron's tower)
  - Extremely powerful
  - Patrolled by Orcs
    - Let one little hobbit in it, and the whole thing could come crashing down
  - OWL

The *S*emantic web

- Ontology as Barad-dur (Sauron's tower)
  - Extremely powerful
  - ~~Patrolled by Orcs~~ Decidable logic basis
    - Let one little ~~hobbit~~ inconsistency in it, and the whole thing could come crashing down
  - OWL

Inconsistency



1537 classes,
1 modeling error
= failure!

The semantic web

- The tower of Babel
  - We will build a tower to reach the sky
  - We only need a little ontological agreement
    - Who cares if we all speak different languages?

This is RDFS

Statistics matter here

Web-scale

Lots of data; finding anything in the mess can be a win

Approaches to data representation

- Objects
- Tables (relational databases)
- Non-relational databases
- Tables (data.frame)
- Graphs

What one can do with semantic web data, now:

People that died in Nazi Germany and if possible, any notable works that they might have created

```
SELECT *
WHERE {
   ?subject dbpprop:deathPlace
<http://dbpedia.org/resource/Nazi_Germany> .
   OPTIONAL {
      ?subject dbpedia-owl:notableworks ?works
   }
}
```

subject	works
:Anne_Frank	:The_Diary_of_a_Young_Girl
:Martin_Bormann	-
:Ir%C3%A8ne_N%C3%A9mirovsky	-
:Erich_Fellgiebel	-
:Friedrich_Ferdinand%2C_Duke_of_Schleswig-Holstein	-
:Friedrich_Olbricht	-
:Ludwig_Beck	-
:Erwin_Rommel	-
:Maurice_Bavaud	-
:Early_Years_of_Adolf_Hitler	-
:Emil_Zegad%C5%82owicz	-
:Friedrich_Fromm	-
:Helmuth_James_Graf_von_Moltk	

SEVENTH FRAMEWORK PROGRAMME

LarKC Large Knowledge Collider

- Scale to the entire web
- Do reasoning with open word assumption
- Retrieval in real-time
- Go beyond logics

- Use cases:
  - Real time city
  - Cancer monographs for WHO
  - Gene expression finding

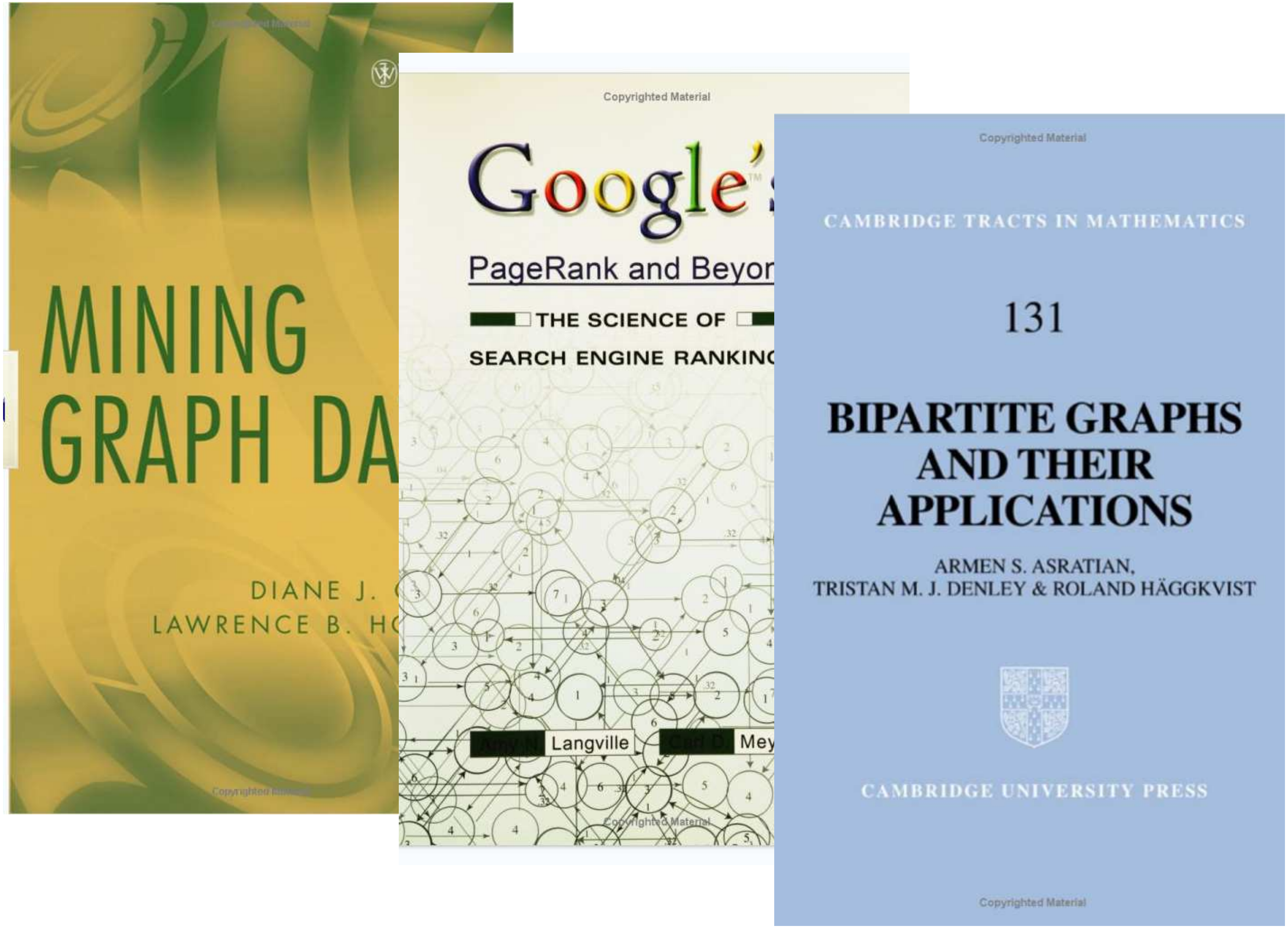
RDF is a graph

- We have lots of interesting statistics that run on graphs
- In many Semantic Web (SW) domains a tremendous amount of statements (expressed as triples) might be true but, in a given domain, **only a small number of statements** is known to be true or **can be inferred to be true**. It thus makes sense to attempt to estimate the truth values of statements by **exploring regularities** in the SW data with machine learning

Scale

- You cannot use the entire thing at once: subsetting

- Are there patterns in knowledge structures that we can use for subsetting?


MINING
GRAPH DA
DIANE J.
LAWRENCE B. H
Copyrighted Material
Google's
PageRank and Beyon
THE SCIENCE OF
SEARCH ENGINE RANKIN
Langville
Mey
Copyrighted Material
CAMBRIDGE TRACTS IN MATHEMATICS
131
BIPARTITE GRAPHS
AND THEIR
APPLICATIONS
ARMEN S. ASRATIAN,
TRISTAN M. J. DENLEY & ROLAND HÄGGKVIST
CAMBRIDGE UNIVERSITY PRESS
Copyrighted Material

Idea

- Graph theory applied to subsetting large graphs
- Developing Semantic Web applications requires handling the RDF data model in a programming language
- Problem: current software is developed in the object-oriented paradigm, programming in RDF is currently triple-based.

Data

IMDB is a big graph:

- 1.4 m movies
- 1.7 m actors
- 11 M connections
  - Movies have votes
- Bipartite network

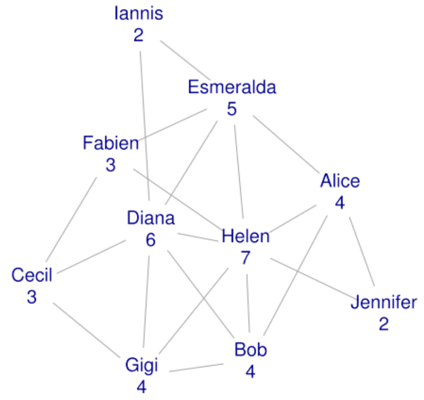
Packages: *igraph:*

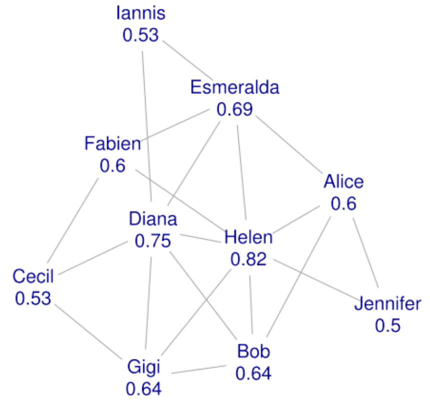
- Nice functions that you cannot find anywhere else
- Uses Sparse Matrices
- Implemented in C
- Some support for bipartite networks

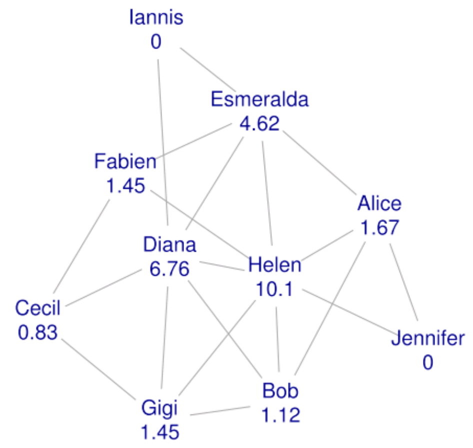
Rmysql, Matrix (sparse m)

Centrality

Method	Formula	Time complexity
Degree	$$C_d(v) = \frac{deg(v)}{n-1}$$ Where $deg(v)$ is the number of connections that v has.	$O(E)$
Closeness	$$C_v = \frac{\lvert v \rvert - 1}{\sum_{i \neq v} d_{vi}}$$ Where d_{vi} is the distance between vertex d and i and $\lvert v \rvert$ is the number of vertices	$O(V^3)$
Betweenness	$$B_v = \sum_{i \neq j, i \neq v, j \neq v} g_{ivj}/g_{ij}$$ Where g_{ivj} is the number of shortest paths between i and j that pass through v, and g_{ij} is the number of paths between i and j that do not go through v	$O(VE)$ time using Brandes' (2001) algorithm; parallelizable (Bader & Madduri, 2006).

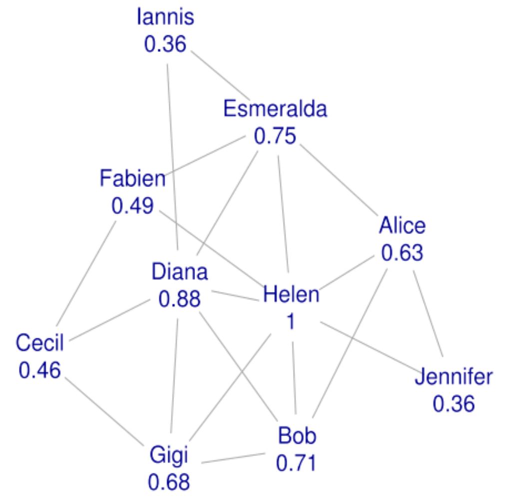


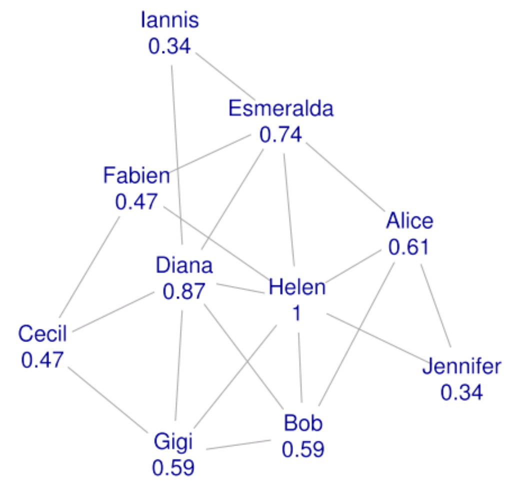




Centrality

Eigenvector $$E_v = \frac{1}{\lambda} A_{iv} E_i, \quad Ax = \lambda x$$ Where A is a matrix that represents a linear transformation, and lambda is the scaling factor (eigenvalue) in the eigenvalue equation (right)	$O(\mathrm{V})$
PageRank $$E_v = \frac{1-d}{\lvert V \rvert} + d \sum_{i=1}^{\lvert V \rvert} A_{iv} E_i$$ d is the damping factor (set at .85 in our experiment, as in the original Brin and Page paper)	$O(\lvert E \rvert \log(1/e))$, e is the prec sion required. The complexity is indepe dent of the number of vertic (Bianchini, 2005, p. 100).

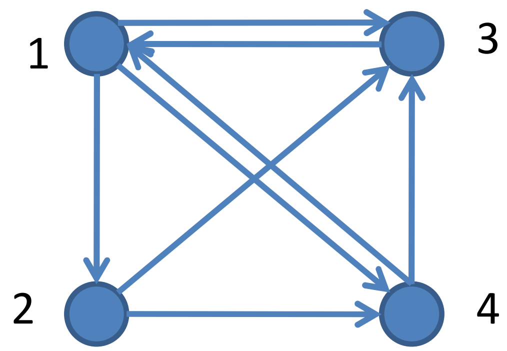




Pagerank

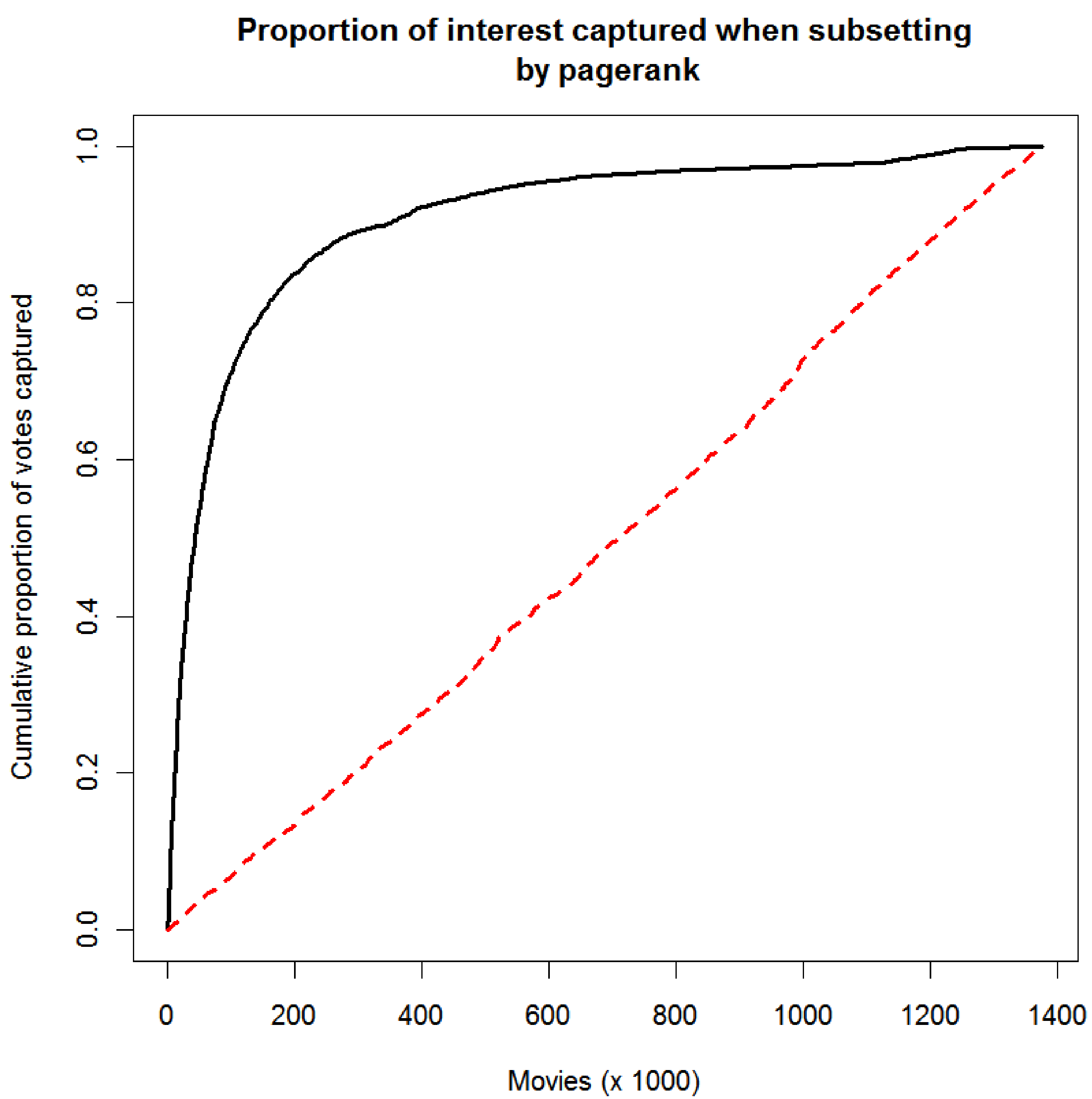
- The pagerank vector is the stationary distribution of a markov chain in a link matrix
- Some assumptions to warrant convergence
- The typical value of d is .85

$$E_v = \frac{1-d}{|V|} + d \sum_{i=1}^{|V|} A_{iv} E_i$$



$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}.$$

```
norm <- function(x) x/sum(x)
norm(eigen(0.15/nVertices + 0.85 * t(A))$vectors[,1])
```


Proportion of interest captured when subsetting
by pagerank
Cumulative proportion of votes captured
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
1200
1400
Movies (x 1000)

Top movies by pageRank in the actor->movie network

degree	pagerank	cluster	imdbID	title	rank	votes
1298	0.000243688252192870	0	822609	Around the World in Eighty Days (1956)	40031	6134
313	0.000103540862390464	0	76352	\Beyond Our Control\" (1968)"	0	0
291	0.000091669 0099912811	0	993780	Gone to Earth (1950)	7.0	291
285	0.000089025 5923652847	0	915626	Deadlands 2: Trapped (2008)	39971	15
424	0.000083882328163772	0	1282574	Stuck on You (2003)	6.0	19709
629	0.000080824 1101098043	0	622100	\Shortland Street\" (1992)"	39850	225

Problems

- Graphs have advantages over RDBMS/tables[1]. But we are used to think in tables
- There is no direct way to handle RDF in R. worth an R package?

ActiveRDF: Object-Oriented Semantic Web Programming

Eyal Oren
eyal.oren@deri.org

Renaud Delbru
renaud.delbru@deri.org

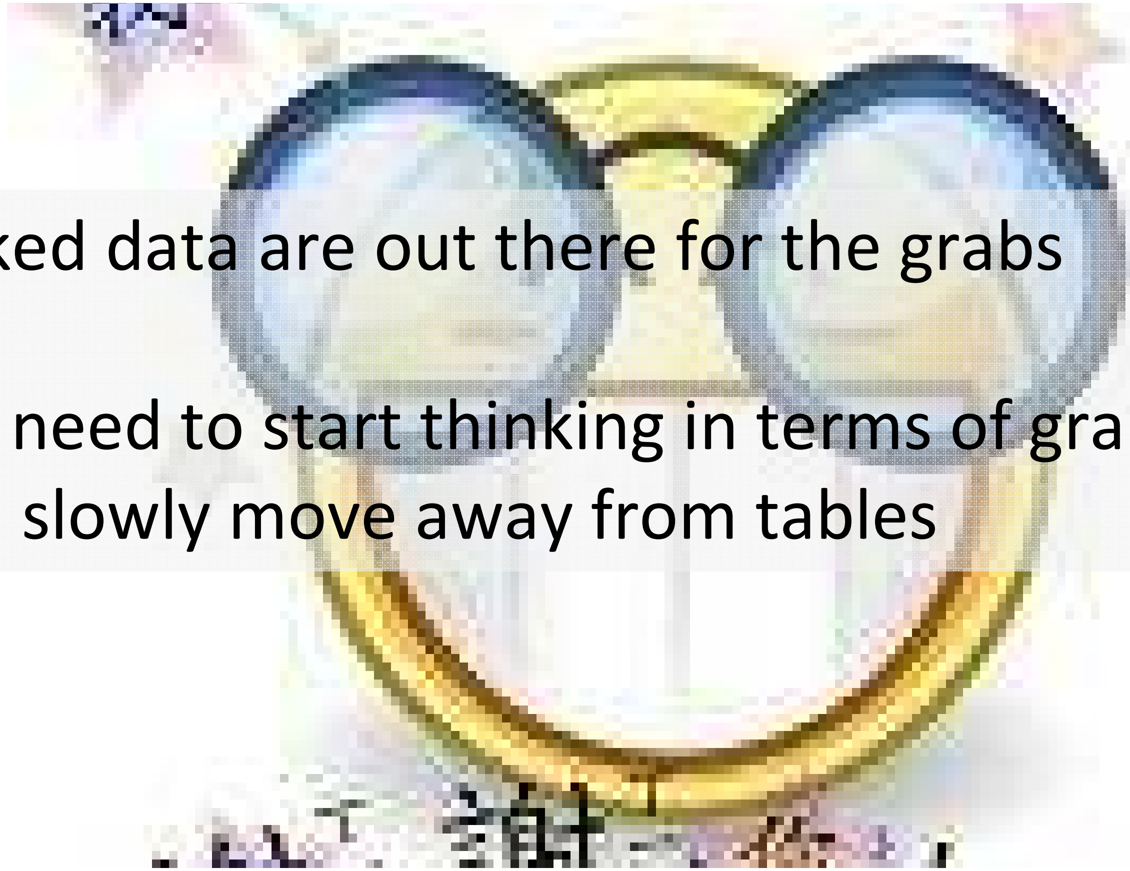
Sebastian Gerke
sebastian.gerke@deri.org

Armin Haller
armin.haller@deri.org

Stefan Decker
stefan.decker@deri.org

Digital Enterprise Research Institute
National University of Ireland, Galway
Galway, Ireland

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.



Linked data are out there for the grabs

We need to start thinking in terms of graphs, and slowly move away from tables

Thanks for your attention

Jose Quesada, quesada@workingcogs.com, http://josequesada.name
Twitter: @Quesada