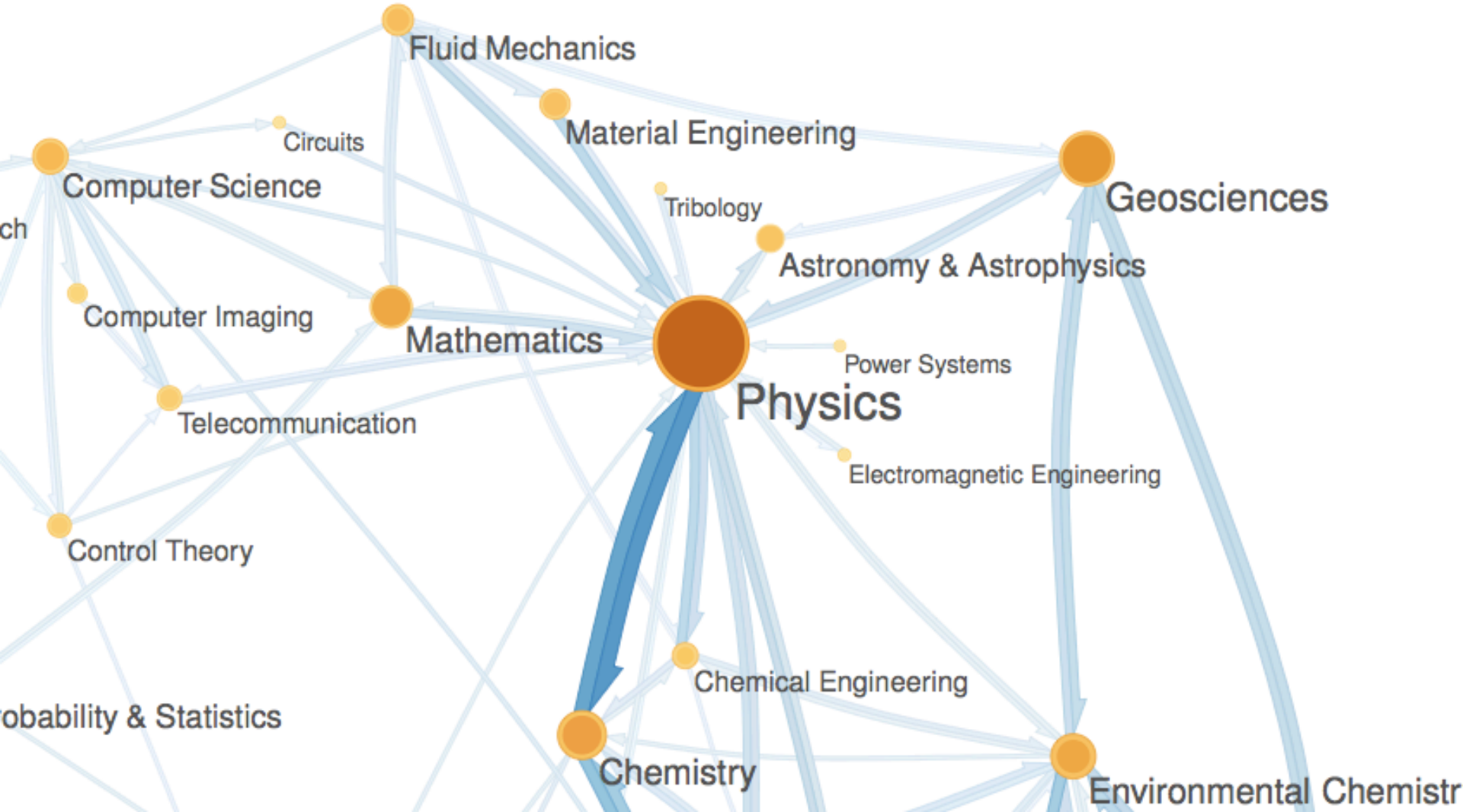
Hierarchically clustering time-directed graphs and the effects of teleportation and memory

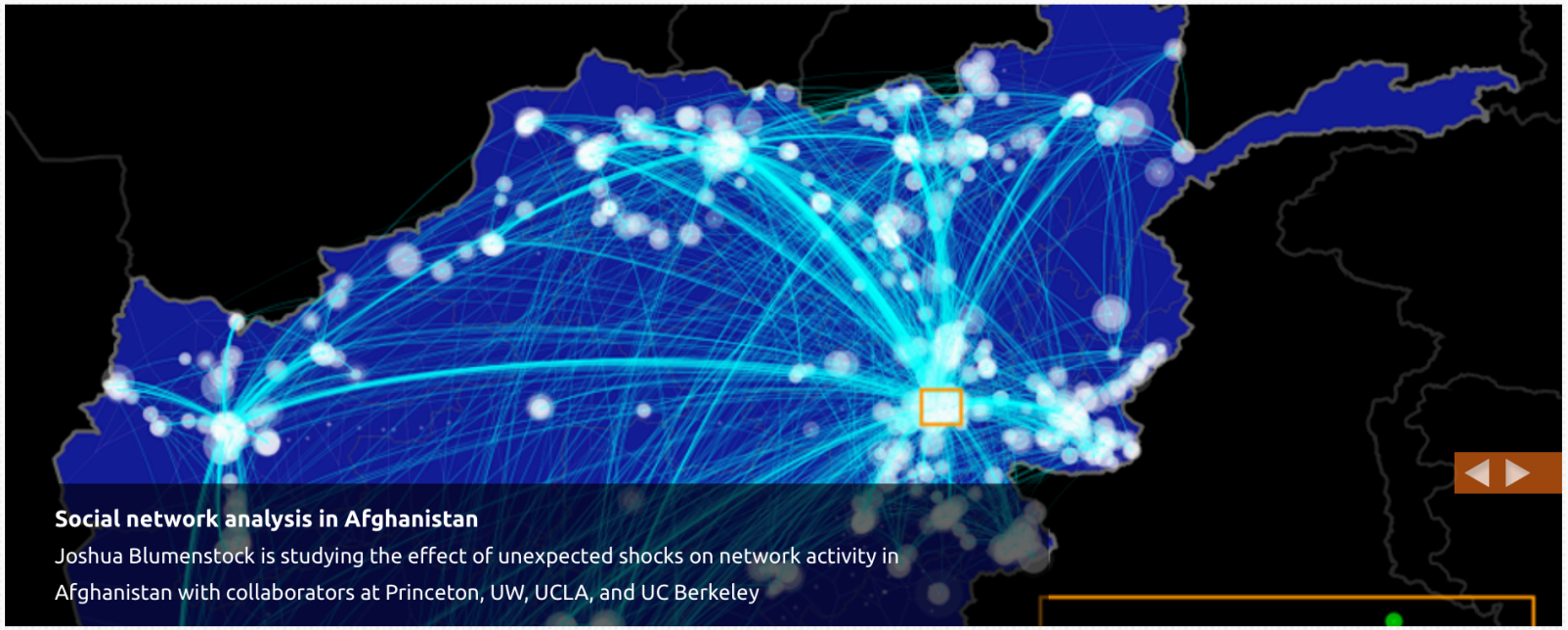
Jevin West, Information School, University of Washington



DataLab

Data Science and Analytics Lab

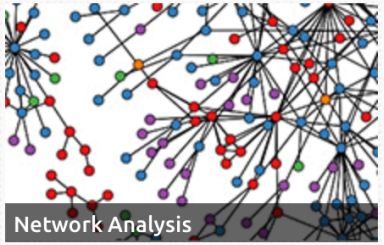
University of Washington Information School



Social network analysis in Afghanistan

Joshua Blumenstock is studying the effect of unexpected shocks on network activity in Afghanistan with collaborators at Princeton, UW, UCLA, and UC Berkeley

Research Focus Areas



News and Updates

28 **Blumenstock at Population Association of America**

What we do

The DataLab is the nexus for research on Data Science and Analytics at the UW iSchool. We study **large-scale, heterogeneous human data** in an

Network Clustering

Graph Partitioning

Community Detection

Block Models

Module Detection

http://www.iloveaba.com/2015/07/no-one-size-does-not-fit-all.html

No one size fits all

- No canonical solution or one generalizable method for all data and all problems (i.e. there is no method that works best on all networks in all situations)
- Need to know the context for why the user is interested in clustering
- We don't even have a definition of a community
- Umbrella term for many facets

Schaub, M.T. et al. (2017) The many facets of community detection in complex networks. *Applied Network Science*

No one size fits all

Cut-based: community detection as minimization of some form of constraint violation

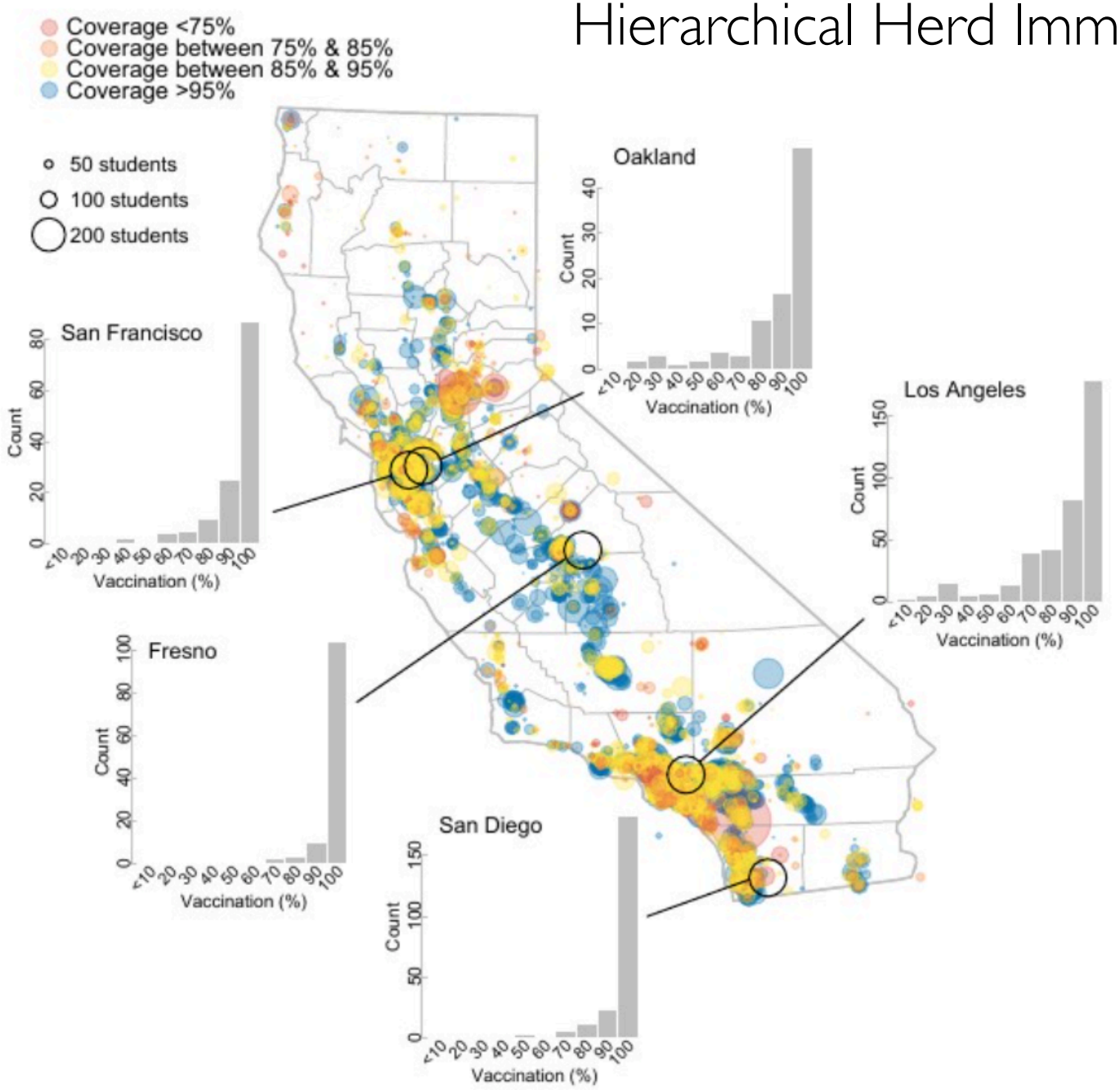
Data clustering: community detection framed as a discretized analogue of data clustering, in which densely knit groups of nodes are to be found

Stochastic equivalence: community detection aiming to identify structurally equivalent nodes in a network, leading to notions such as stochastic block models

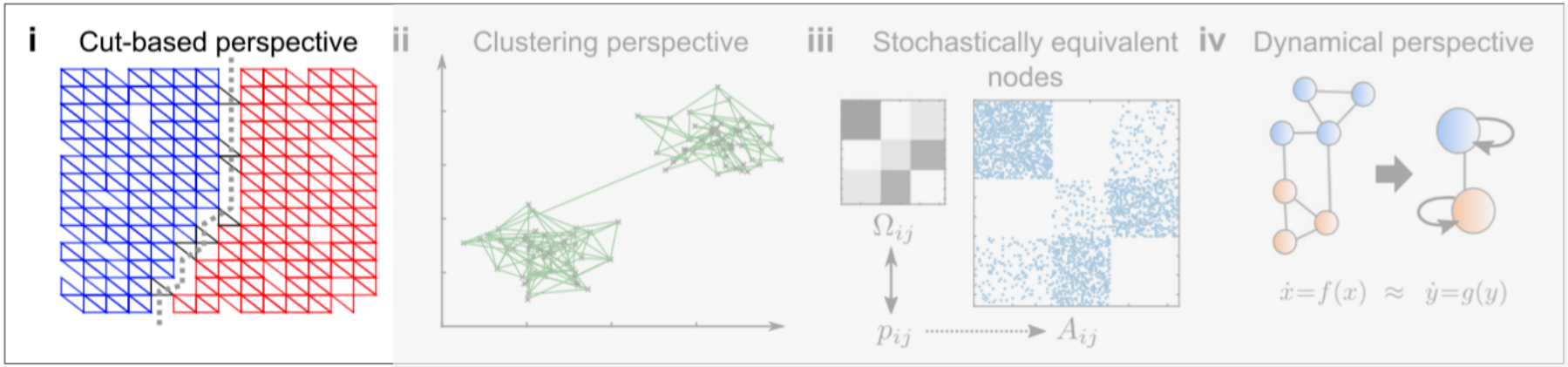
Dynamics perspective: community detection looking for simplified descriptions of the dynamical flows occurring on the network, that is, some form of dynamical model reduction

Schaub, M.T. et al. (2017) The many facets of community detection in complex networks. *Applied Network Science*

Hierarchical Herd Immunity



Community Detection Perspectives



Circuit layout
Minimizing cuts
Load balancing
Eigenvectors
Spectral methods
Image segmentation

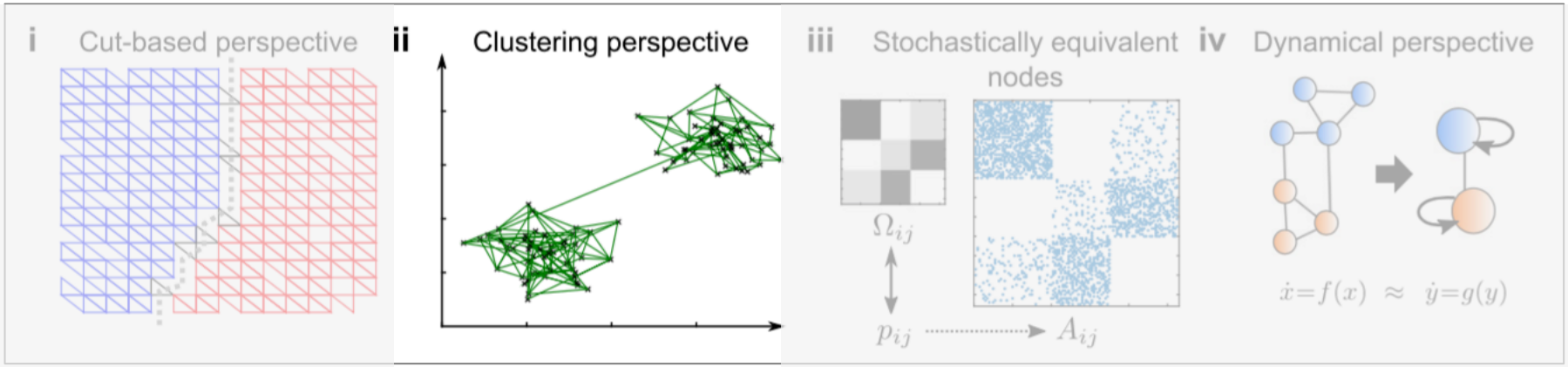
Data Clustering
Maximizing node density
unknown *k*, unbalanced
Conductance
Local, global
Modularity

Social Networks
Connectivity Profiles
Stochastic equivalence
SBMs, LFR
p-values, hypothesis testing
Bipartite treatment
Predict missing links

System behavior, processes
Non-adjacency focused
Airline network
Markovian diffusion process
Undirected, Directed
InfoMap

Schaub, M.T. et al. (2017) The many facets of community detection in complex networks. *Applied Network Science*

Community Detection Perspectives



Circuit layout
Minimizing cuts
Load balancing
Eigenvectors
Spectral methods
Image segmentation

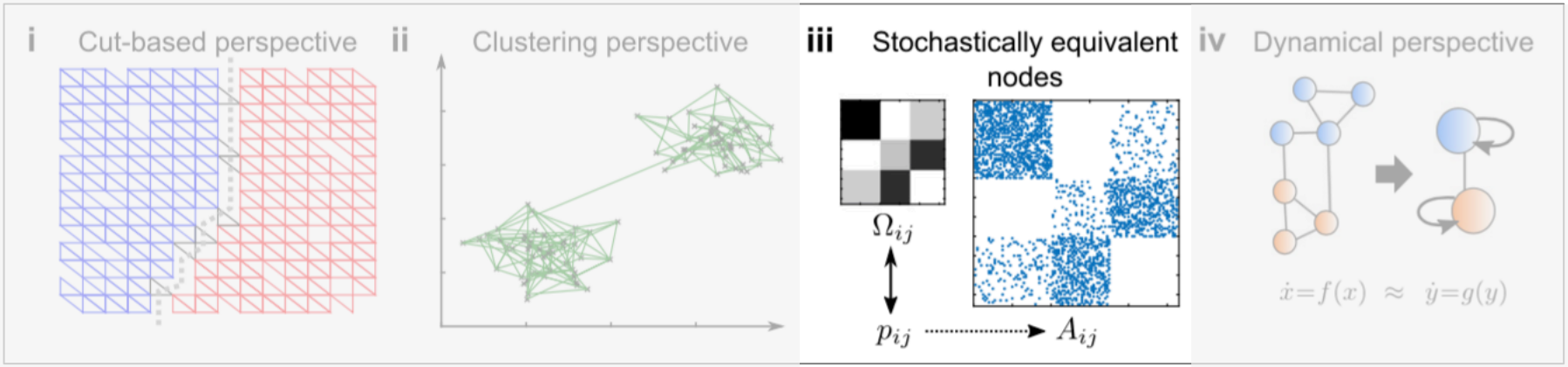
Data Clustering
Maximizing node density
unknown *k*, unbalanced
Conductance
Local, global
Modularity

Social Networks
Connectivity Profiles
Stochastic equivalence
SBMs, LFR
p-values, hypothesis testing
Bipartite treatment
Predict missing links

System behavior, processes
Non-adjacency focused
Airline network
Markovian diffusion process
Undirected, Directed
InfoMap

Schaub, M.T. et al. (2017) The many facets of community detection in complex networks. *Applied Network Science*

Community Detection Perspectives



Circuit layout
Minimizing cuts
Load balancing
Eigenvectors
Spectral methods
Image segmentation

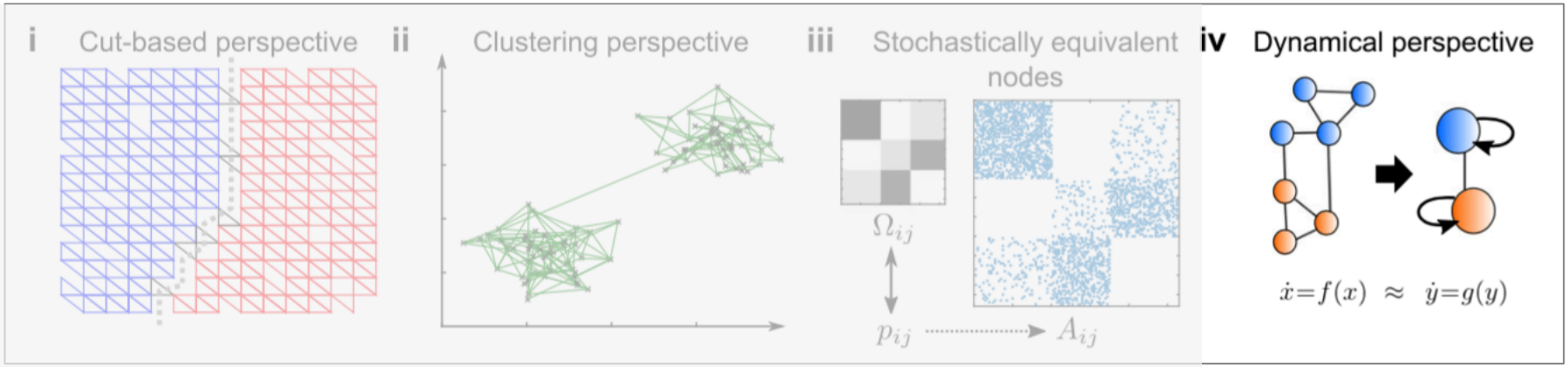
Data Clustering
Maximizing node density
unknown *k*, unbalanced
Conductance
Local, global
Modularity

Social Networks
Connectivity Profiles
Stochastic equivalence
SBMs, LFR
p-values, hypothesis testing
Bipartite treatment
Predict missing links

System behavior, processes
Non-adjacency focused
Airline network
Markovian diffusion process
Undirected, Directed
InfoMap

Schaub, M.T. et al. (2017) The many facets of community detection in complex networks. *Applied Network Science*

Community Detection Perspectives



Circuit layout
Minimizing cuts
Load balancing
Eigenvectors
Spectral methods
Image segmentation

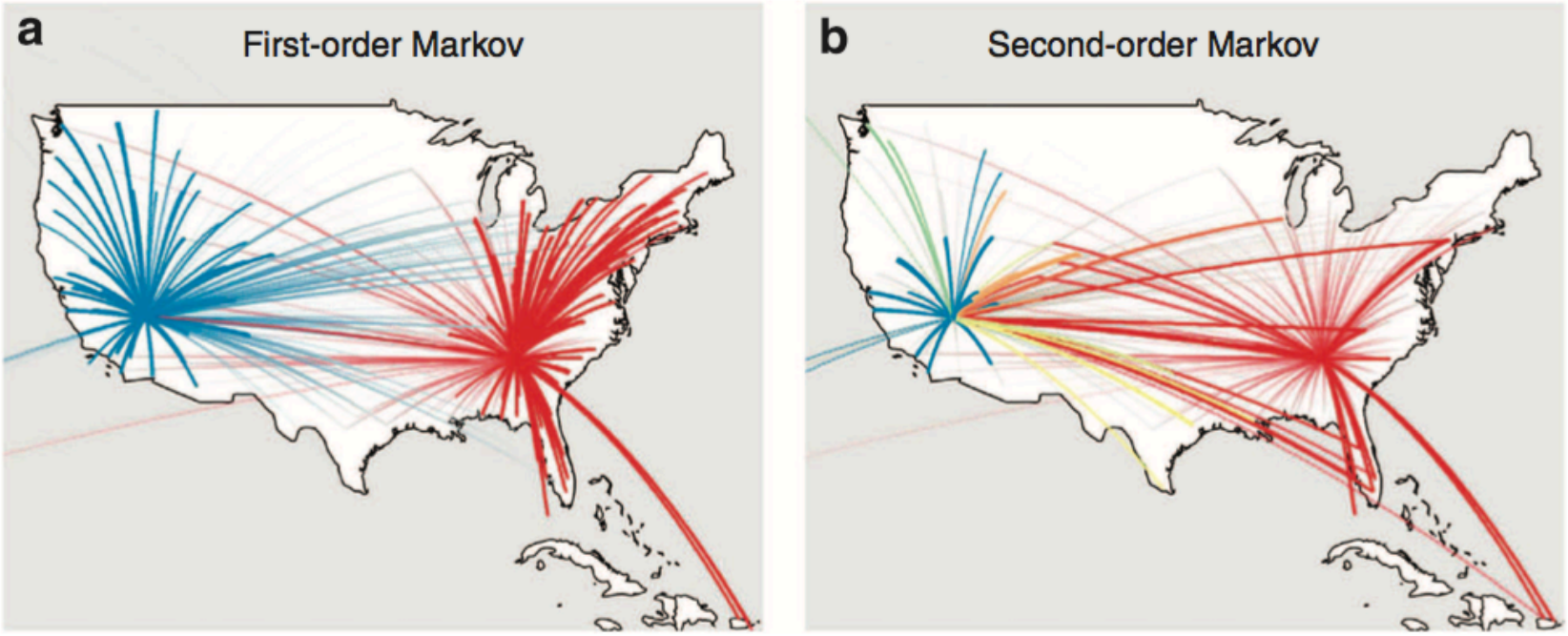
Data Clustering
Maximizing node density
unknown *k*, unbalanced
Conductance
Local, global
Modularity

Social Networks
Connectivity Profiles
Stochastic equivalence
SBMs, LFR
p-values, hypothesis testing
Bipartite treatment
Predict missing links

System behavior, processes
Non-adjacency focused
Airline network
Markovian diffusion process
Undirected, Directed
InfoMap

Schaub, M.T. et al. (2017) The many facets of community detection in complex networks. *Applied Network Science*

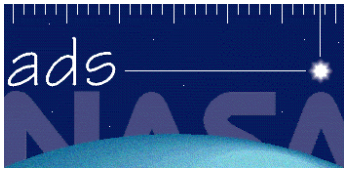
Higher Resolution Maps



Rosvall et al. (2014) Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*

In the spirit of clustering context…

The Scholarly Graph



The Scholarly Graph

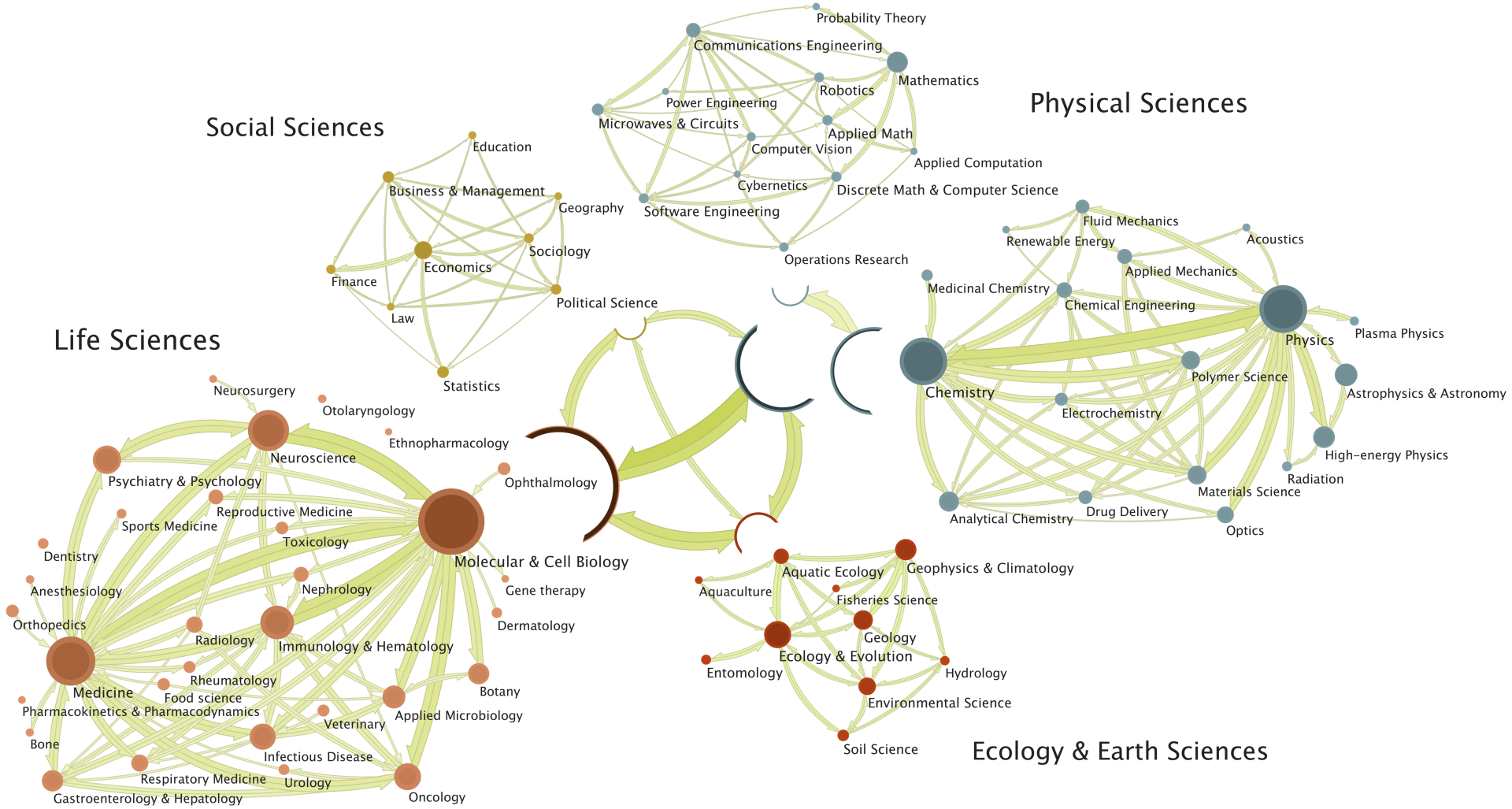
Tens of millions articles, patents, books

Billions of citation links

Years: 1600 – 2016

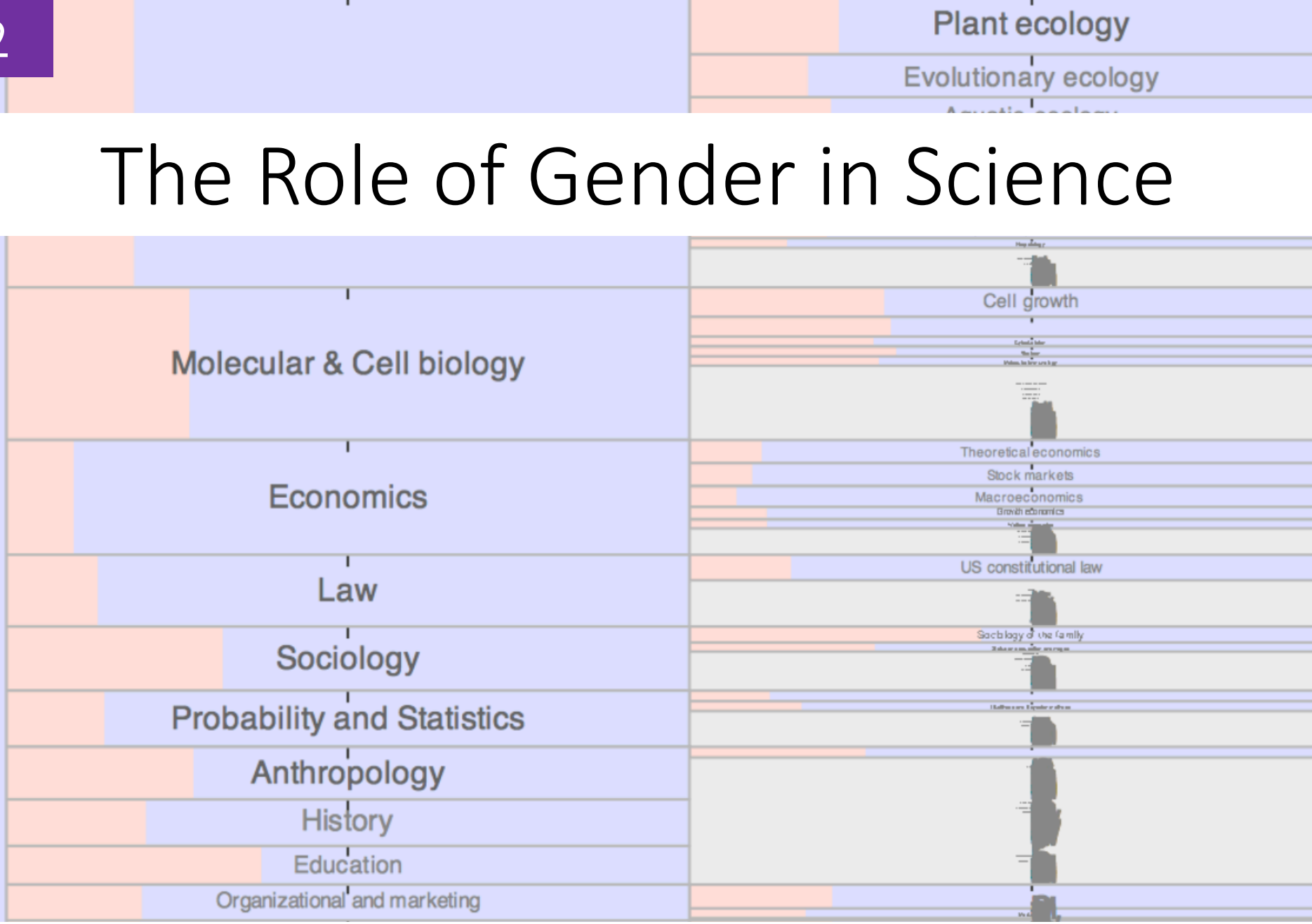
1. Mapping Knowledge Domains
2. Science of Science
3. Hierarchical Navigation
4. Recommendation

Mapping Knowledge Domains



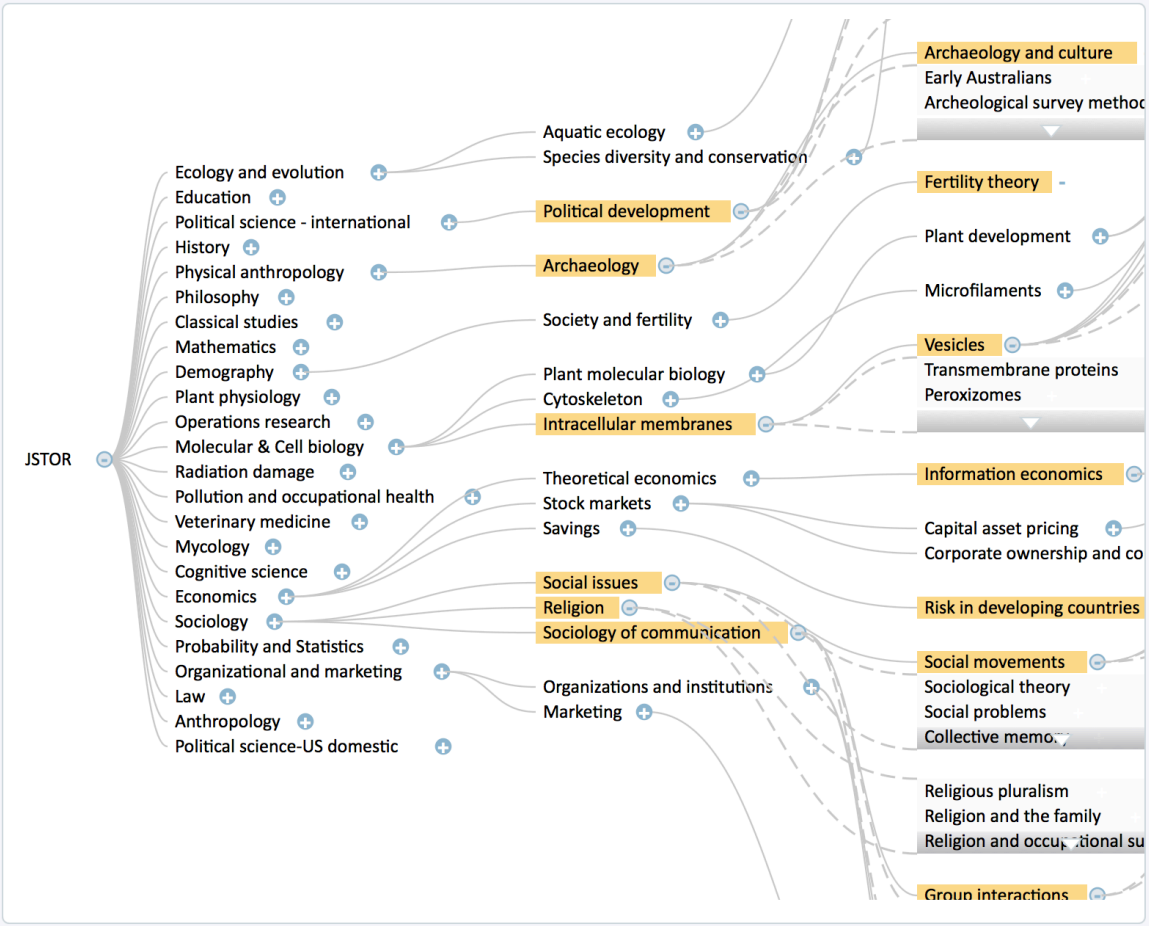
Rosvall, Martin, and Carl T. Bergstrom. "Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems." PloS one 6.4 (2011): e18209.

The Role of Gender in Science



West, J.D. (2012) The Role of Gender in Scholarly Authorship. *PLoS One*

Hierarchical Navigation



Find Papers

by **title** network

by **field**

by **author**

by **journal**

Active Queries: clear all

keyword: **network**

Top Papers Sort by Year (newest)

Using Siting Algorithms in the Design of Marine Reserve Networks
Heather Leslie - *Ecological Applications* (2003)

Mechanism of Filopodia Initiation by Reorganization of a Dendritic Network
Tatyana Svitkina - *The Journal of Cell Biology* (2003)

Network Structure and Knowledge Transfer: The Effects of Cohesion and Range
Ray Reagans - *Administrative Science Quarterly* (2003)

A General Model for Designing Networks of Marine Reserves
Enric Sala - *Science* (2002)

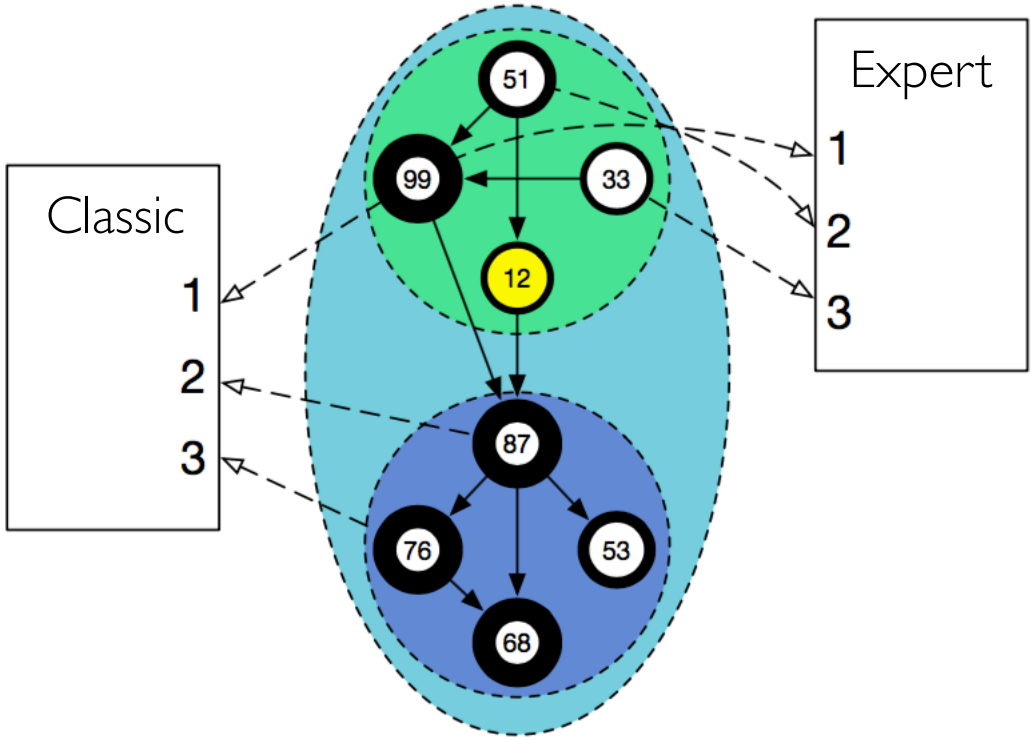
The Density of Social Networks and Fertility Decisions: Evidence from South Nyanza District, Kenya
Hans-Peter Kohler - *Demography* (2001)

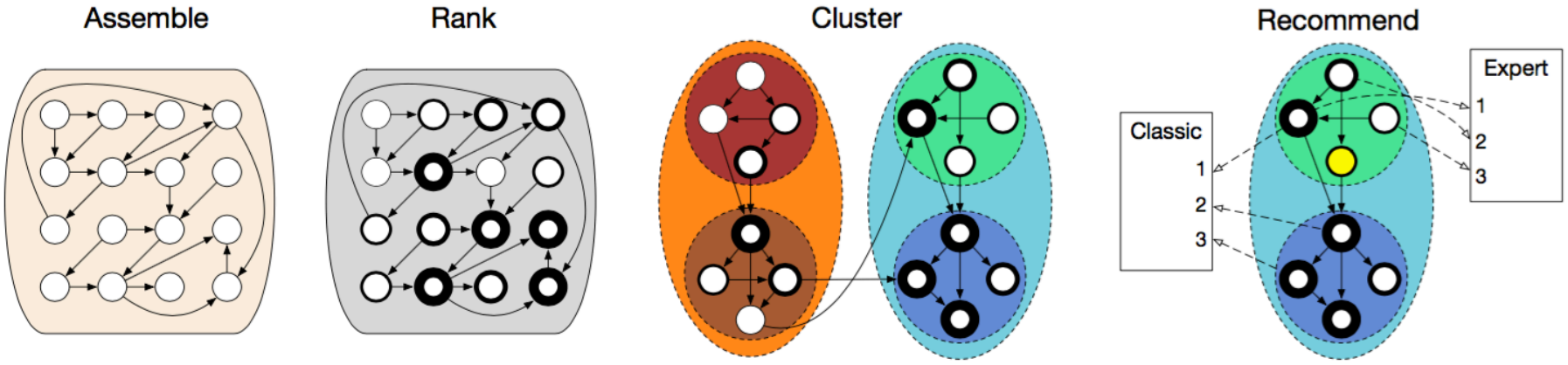
A New Dynamin-Like Protein, ADL6, Is Involved in Trafficking from the trans-Golgi Network to the Central Vacuole in Arabidopsis
Jing Bo Jin - *The Plant Cell* (2001)

Comparing Sequenced Segments of the Tomato and Arabidopsis Genomes: Large-Scale Duplication Followed by Selective Gene Loss Creates a Network of Synteny
Hsin-Mei Ku - *Proceedings of the National Academy of Sciences of the United States of America* (2000)

A Noncooperative Model of Network Formation
Venkatesh Bala - *Econometrica* (2000)

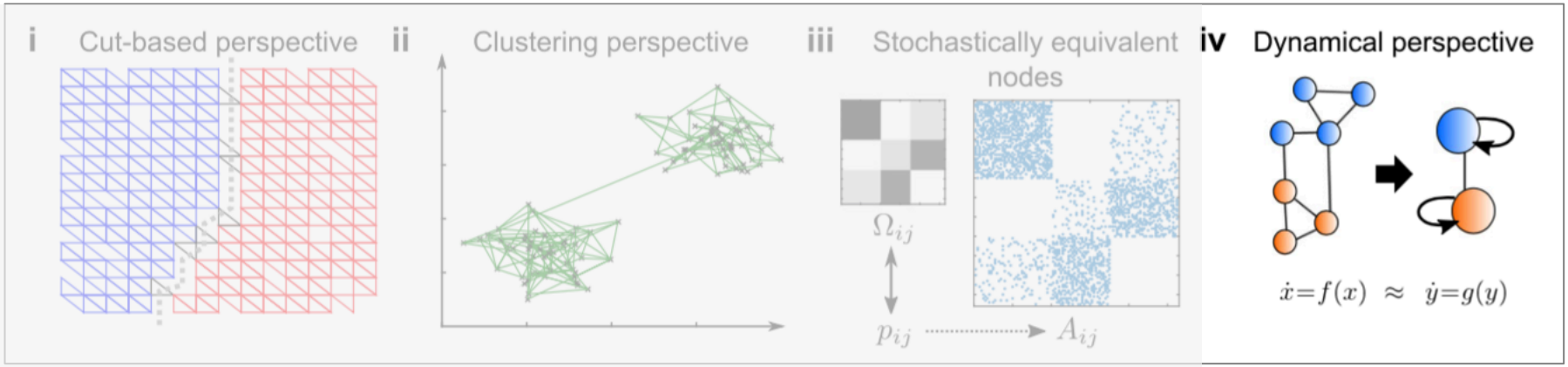
Recommendation





West, Wesley-Smith, Bergstrom (2016) A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE, Transactions on Big Data* (in press)

Community Detection Perspectives



Circuit layout
Minimizing cuts
Load balancing
Eigenvectors
Spectral methods
Image segmentation

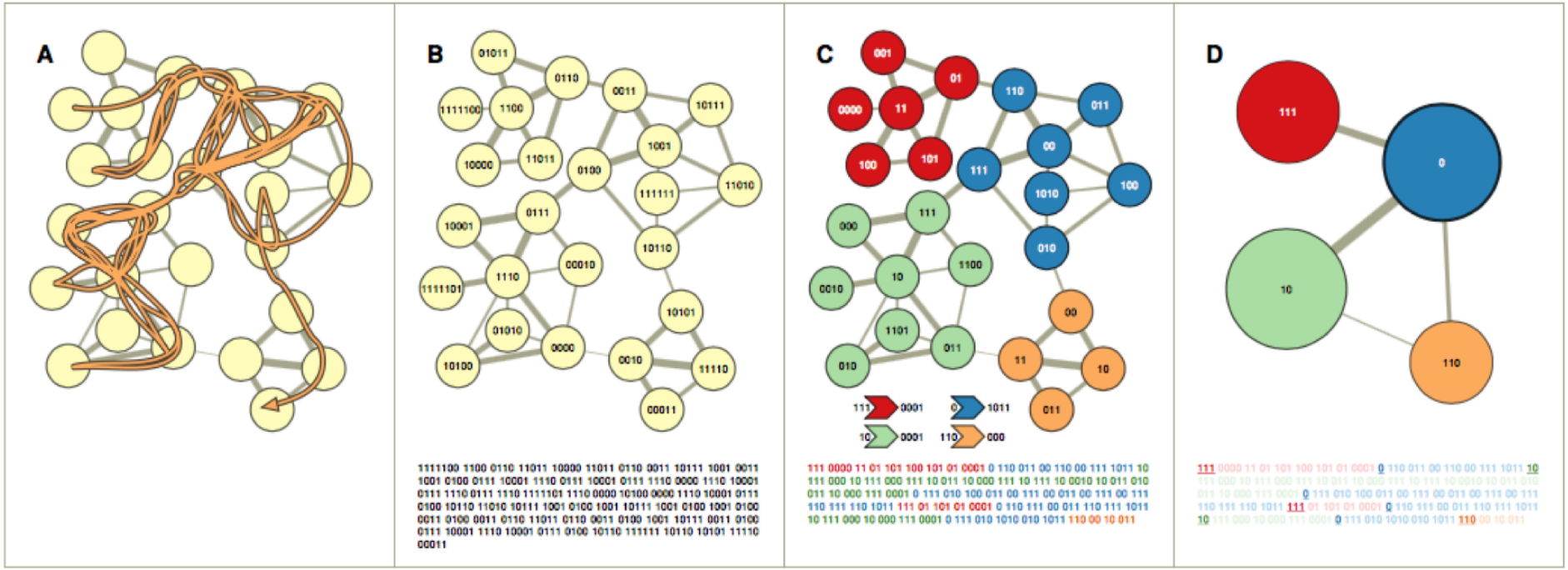
Data Clustering
Maximizing node density
unknown *k*, unbalanced
Conductance
Local, global
Modularity

Social Networks
Connectivity Profiles
Stochastic equivalence
SBMs, LFR
p-values, hypothesis testing
Bipartite treatment
Predict missing links

System behavior, processes
Non-adjacency focused
Airline network
Markovian diffusion process
Undirected, Directed
InfoMap

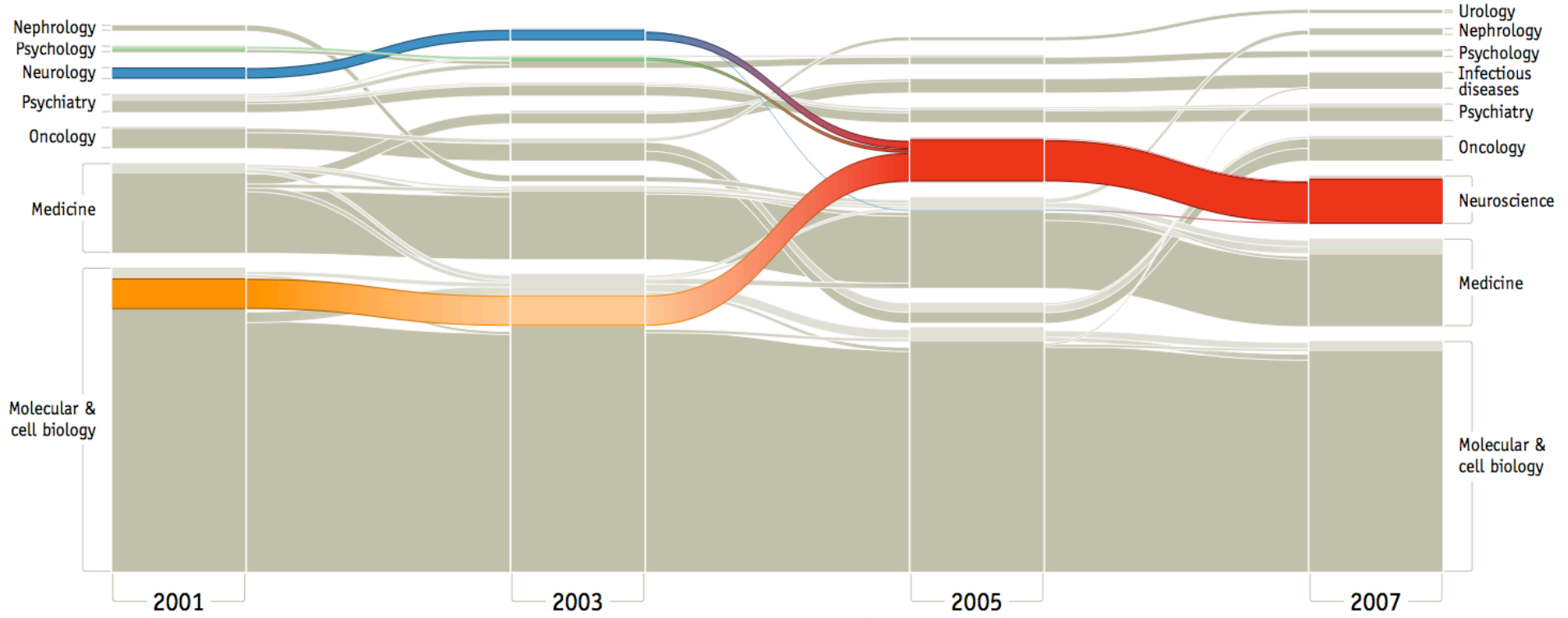
Schaub, M.T. et al. (2017) The many facets of community detection in complex networks. *Applied Network Science*

Finding regularities in citation networks



Rosvall and Bergstrom (2008) PNAS

The Emergence of Neuroscience



Rosvall and Bergstrom (2010) PLoS One

Data

Compressing ⟷ Finding patterns

If we can find a good code for describing flow on a network, we will have solved the dual problem of finding the important structures with respect to that flow.

The map equation

frequency of inter-module movements

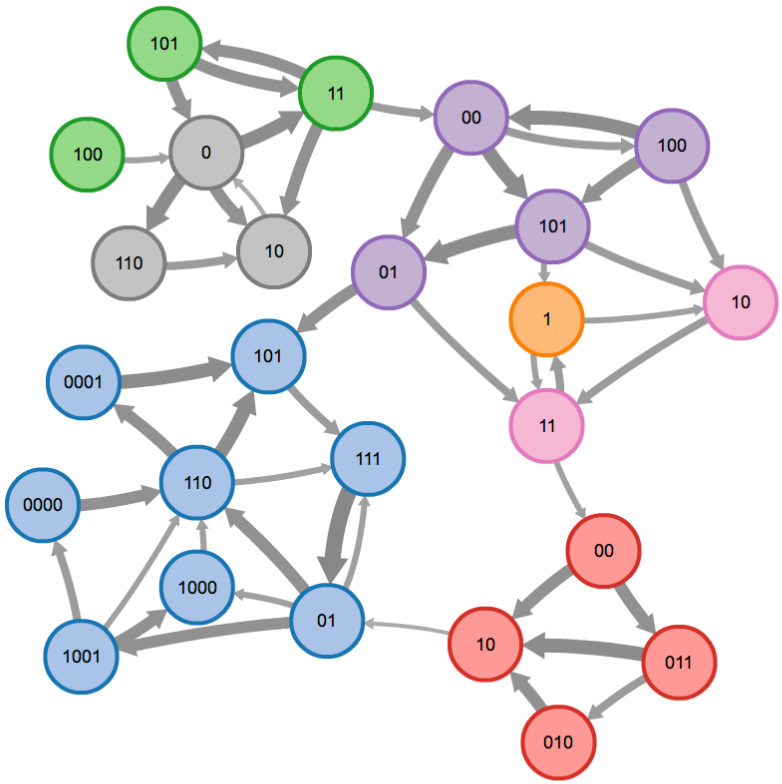
frequency of movements within module *i*

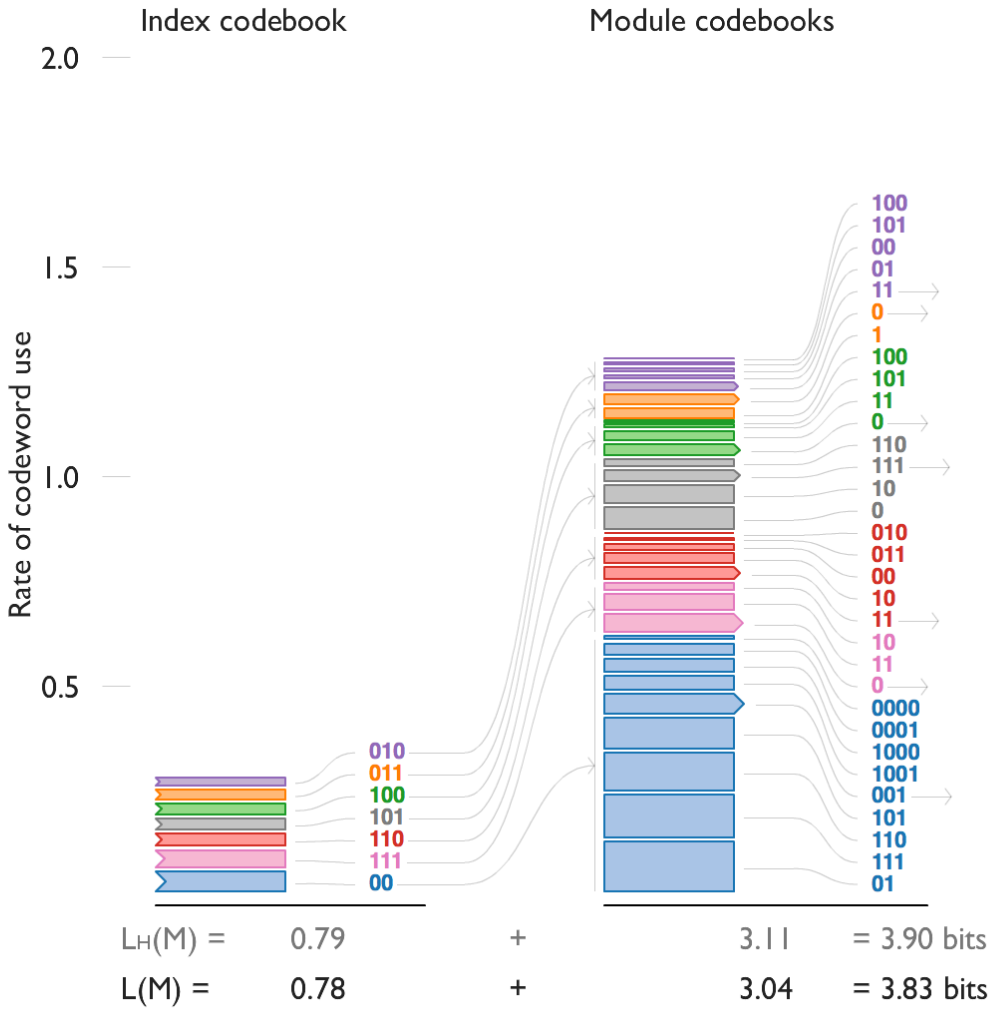
$$L(\mathbf{M}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^i_{\circlearrowright} H(\mathcal{P}^i)$$

code length of module names

code length of node names in module *i*

Rosvall and Bergstrom (2008) *PNAS*

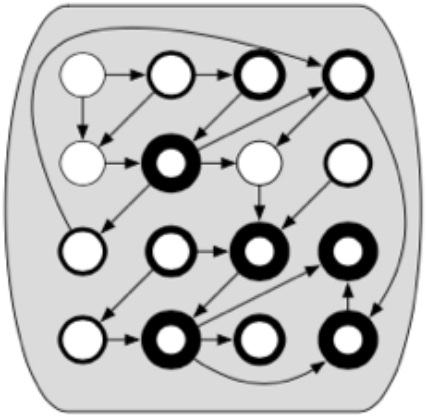




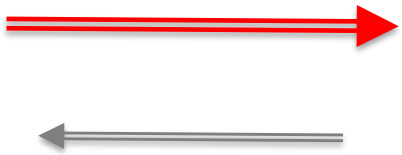
Mapequation.org, Daniel Edler

The relationship between ranking and clustering

Ranking



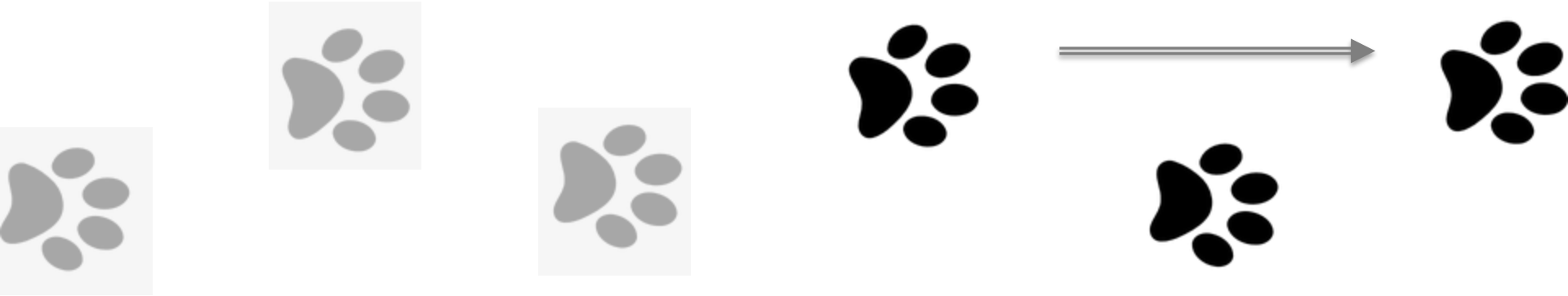
Dynamics



Clustering

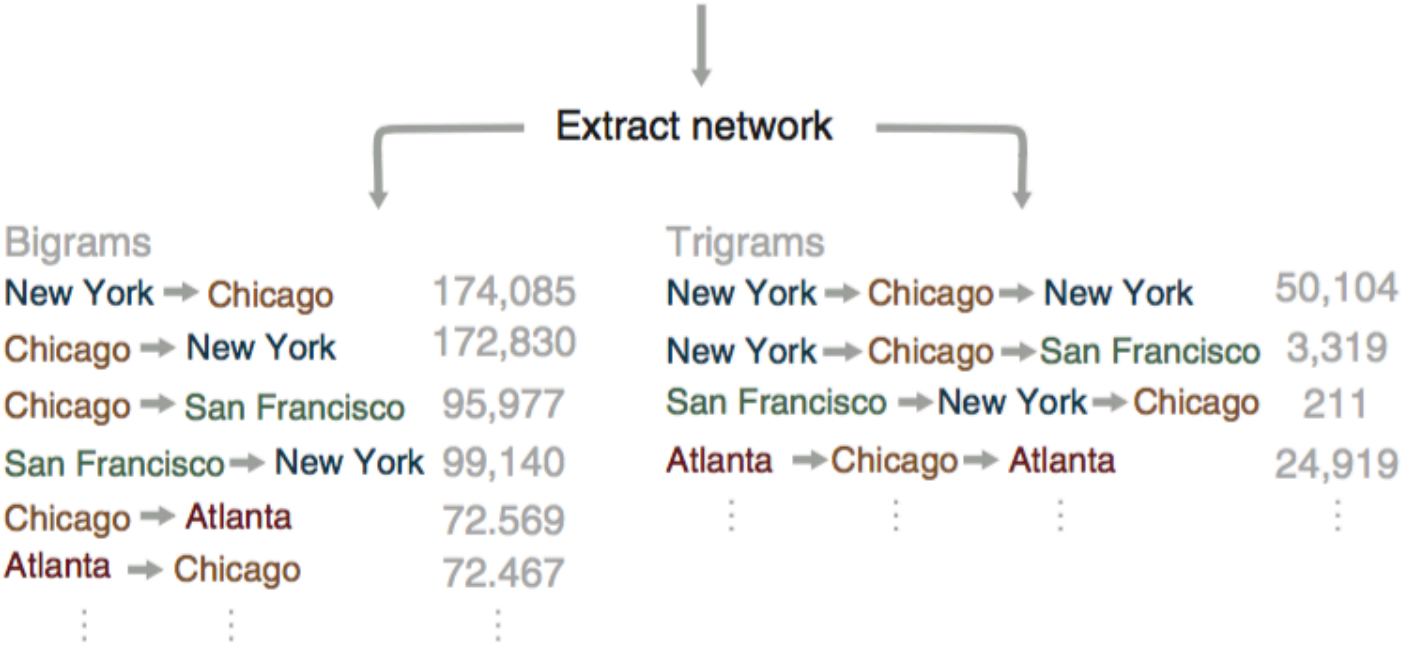
Structure

Step Length, Teleportation and Memory



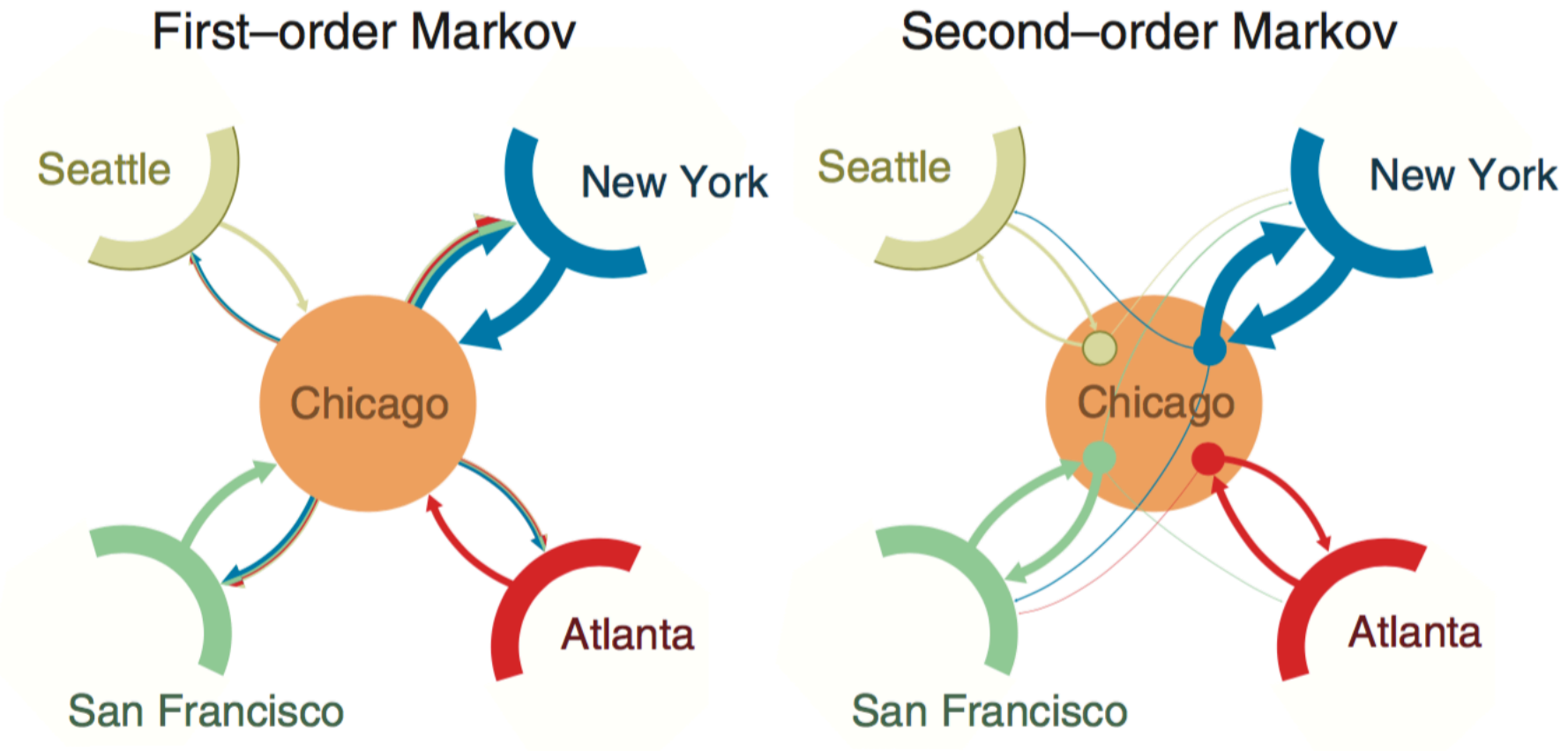
..and their effects on ranking and clustering

Memory: capturing higher order dynamics



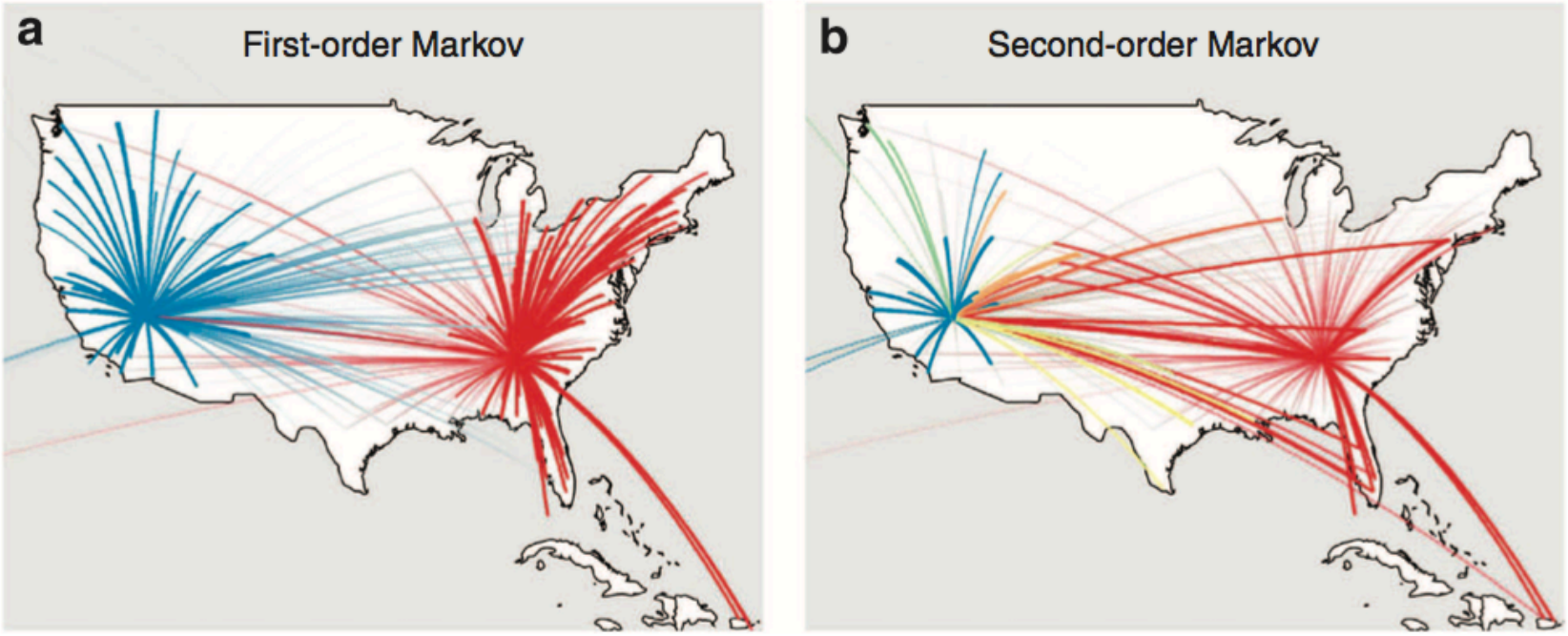
Rosvall et al. (2014) Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*

Memory: capturing higher order dynamics



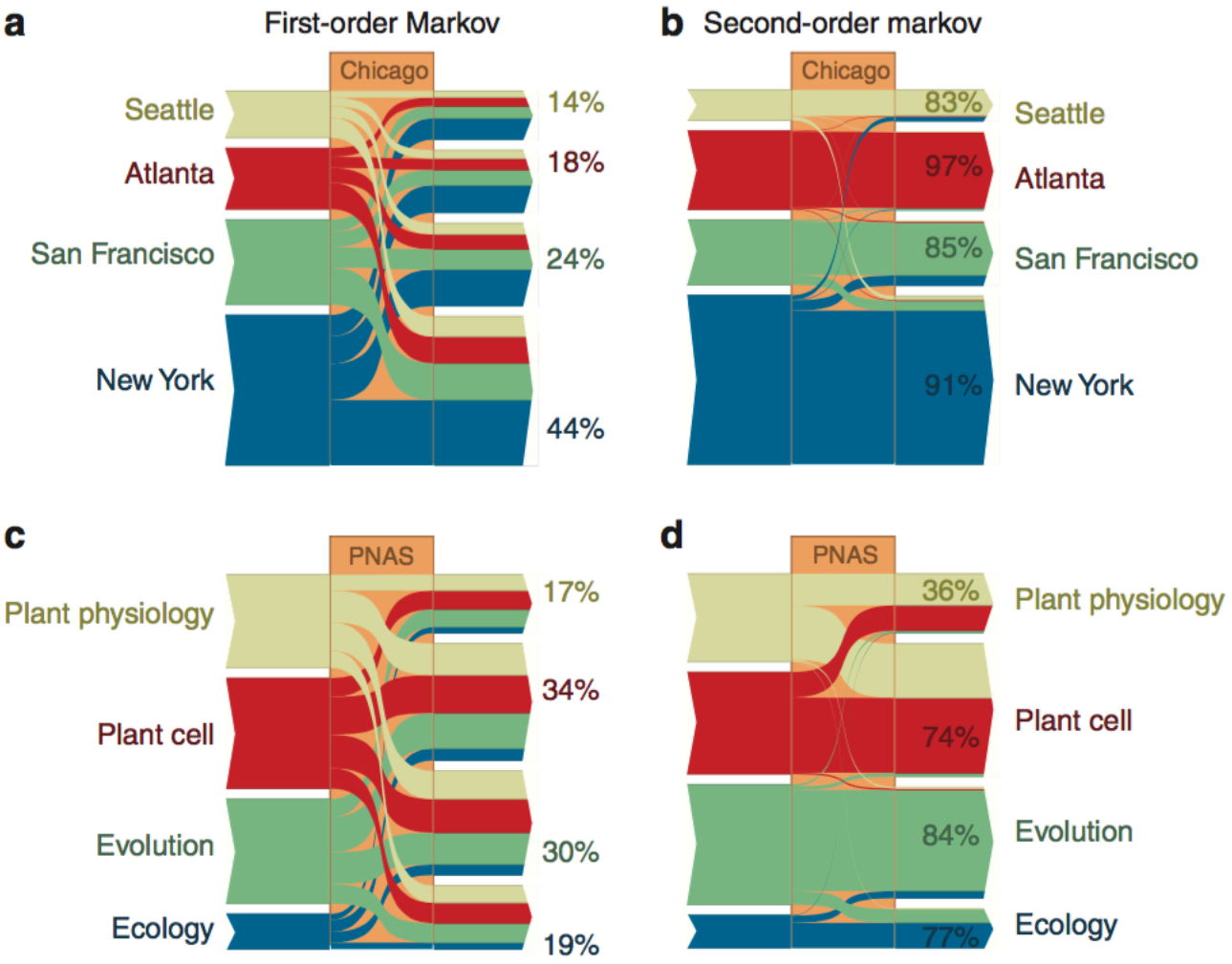
Rosvall et al. (2014) Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*

Higher Resolution Maps



Rosvall et al. (2014) Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*

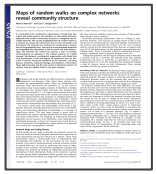
Higher Order Dynamics



Rosvall et al. (2014) Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*

Citation Networks Types

Journal-Level Networks
(Memory)



Article-level Networks

Time-Directed (Acyclic) Graphs

PageRank Variants (EigenFactor)

$$P = \alpha H + (1 - \alpha)\, a.e^{T}$$

Matrix representing the random walk over citations

Probability of not teleporting

Cross-citation Matrix dictating the structure of the citation network

Probability of teleporting to completely new journal weighted by the number of articles in that journal

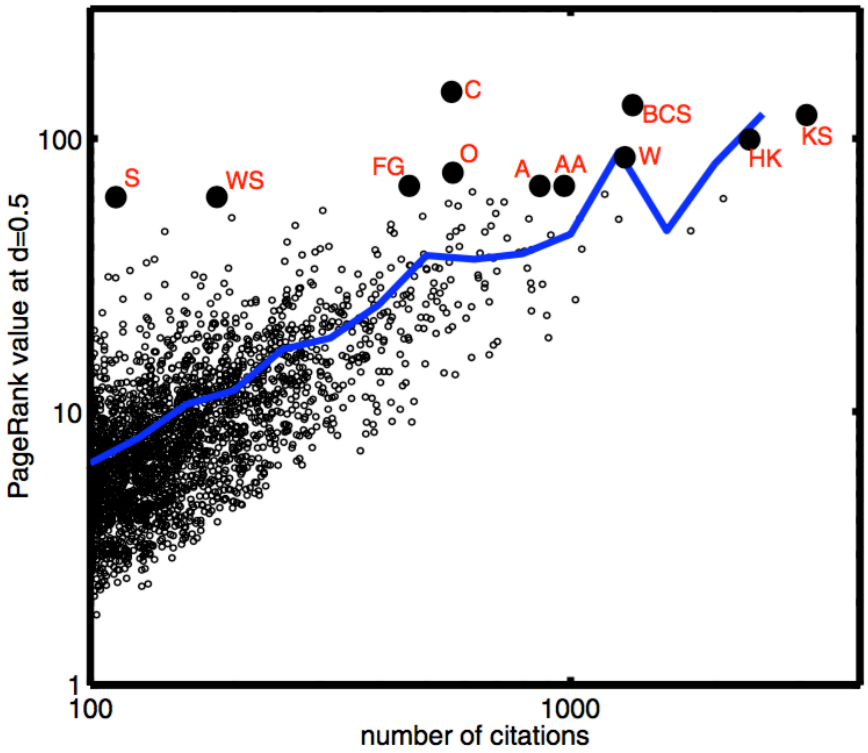
$$EF = 100 \frac{H\pi}{\sum_i [H\pi]_i}$$

Leading eigenvector of the random walk matrix P.

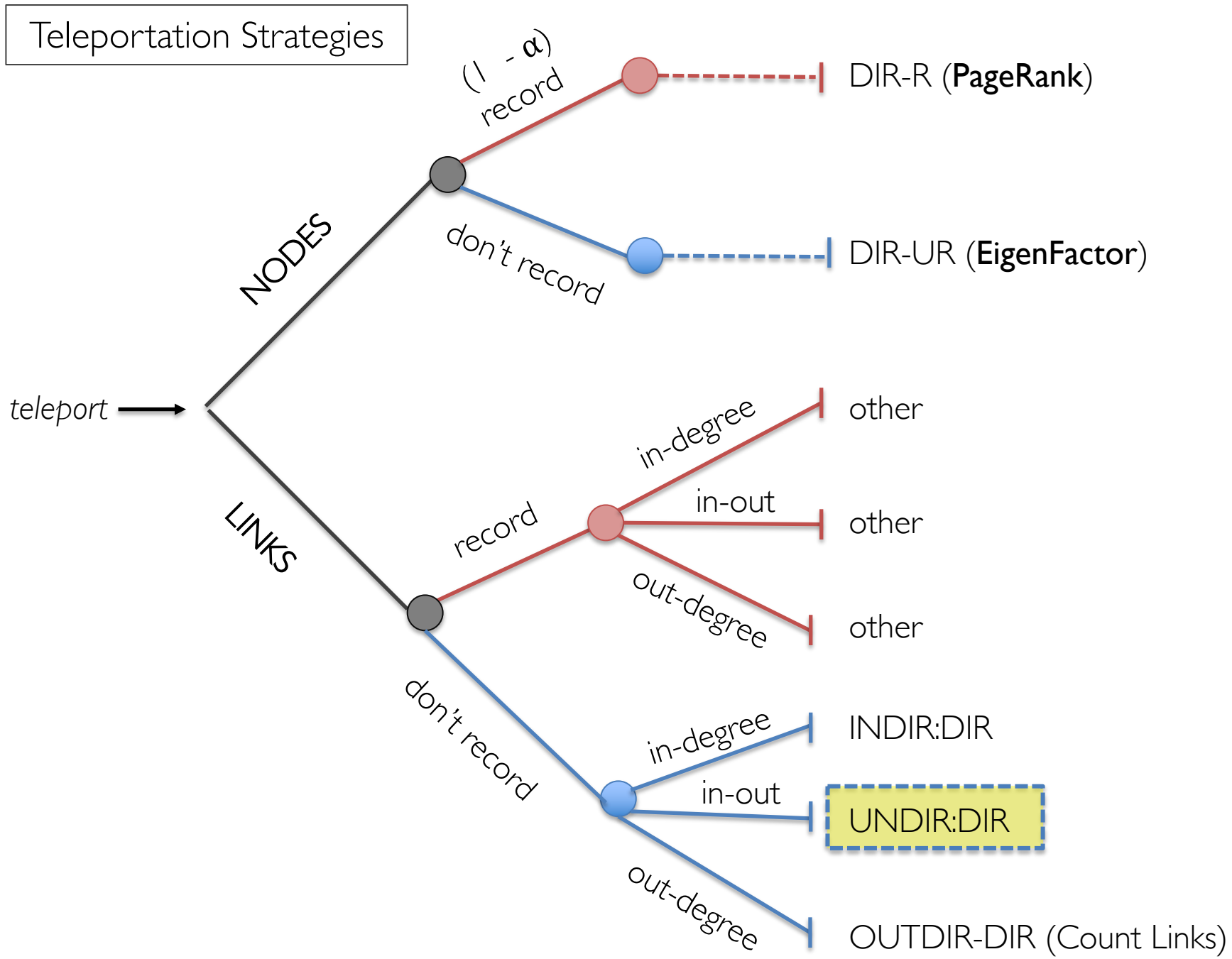
Normalization

West, JD et al. (2010) *College of Research Libraries*

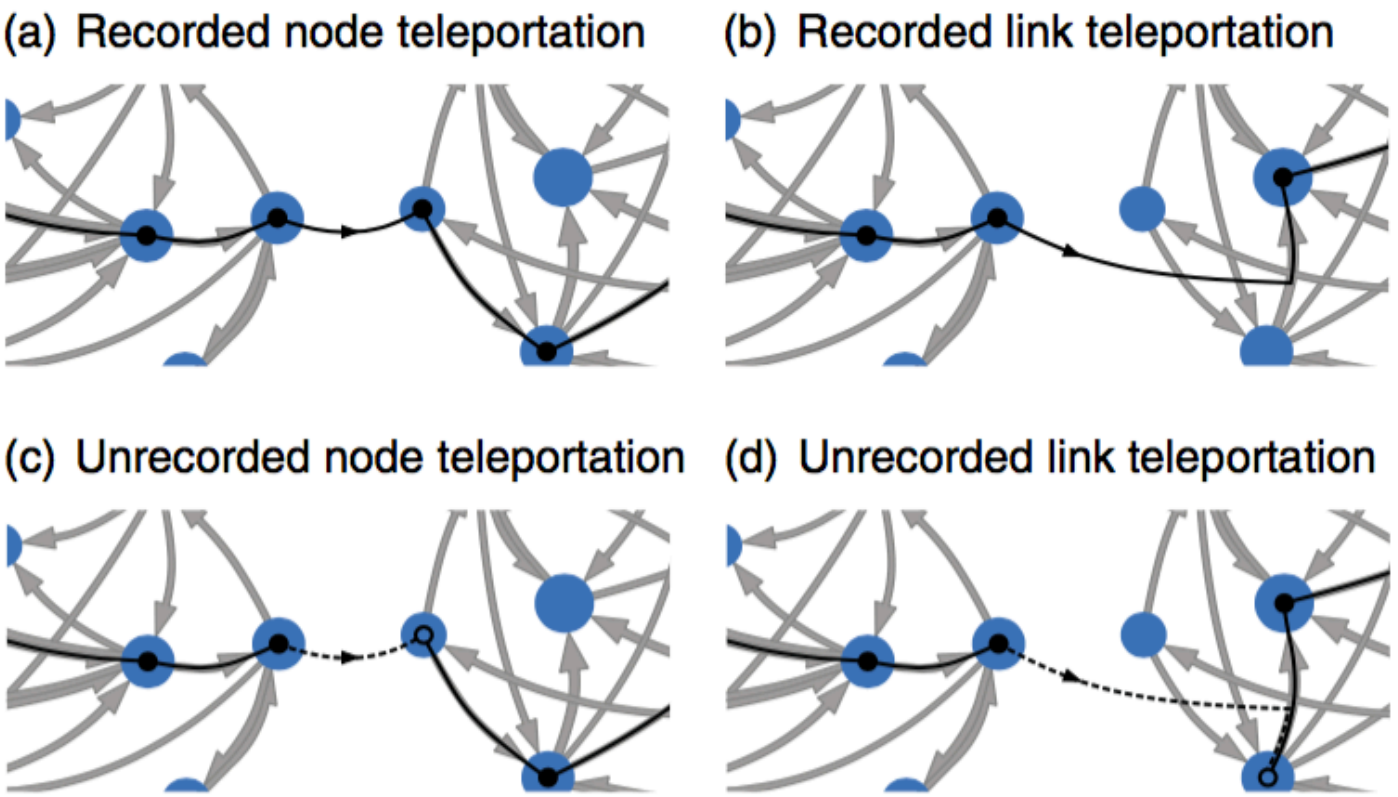
PageRank Pitfalls



Maslov, S. & Redner, S. (2008) Promise and Pitfalls of Extending Google's PageRank Algorithm to Citation Networks. *The Journal of Neuroscience*

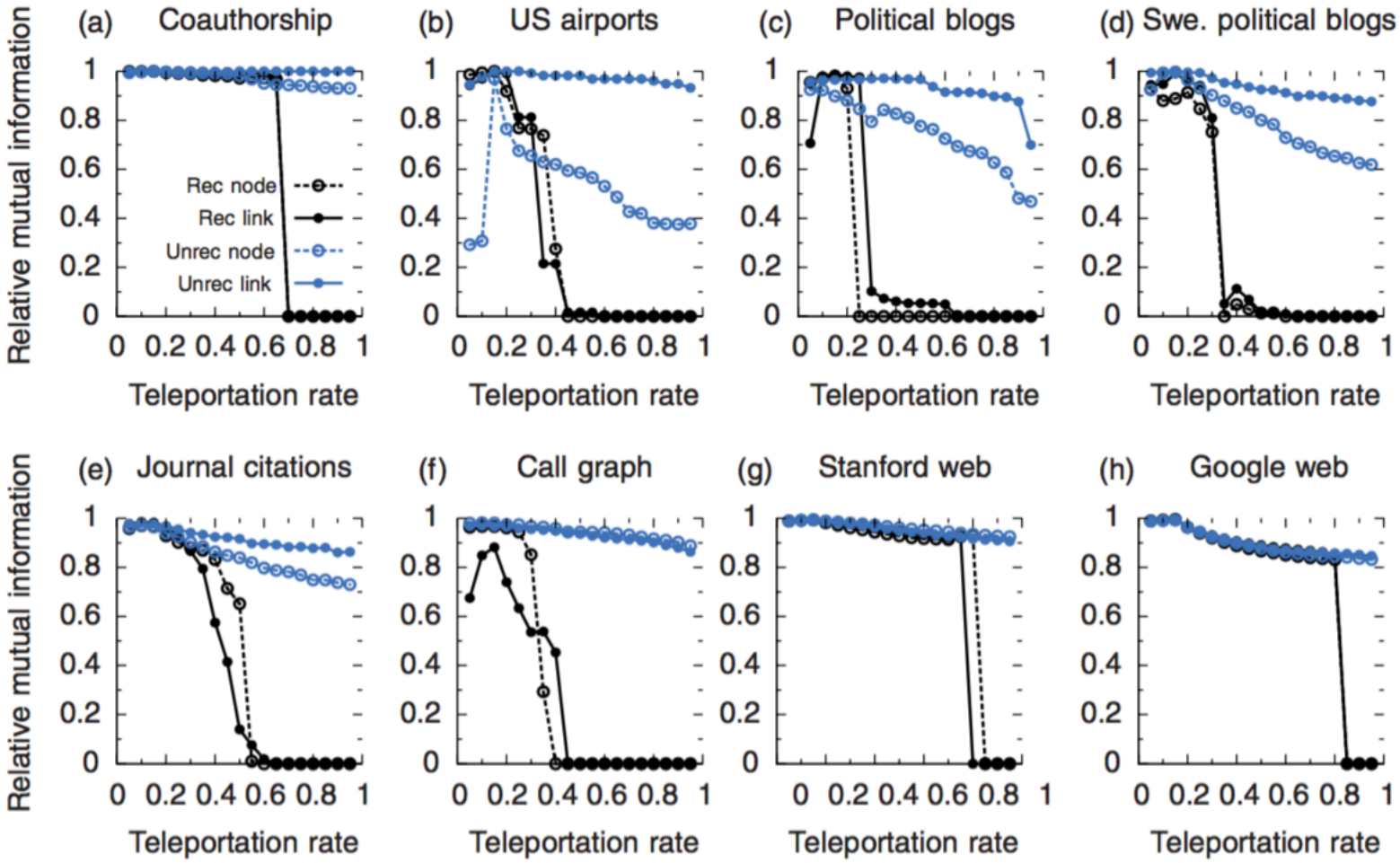

Teleportation Strategies
teleport
NODES
LINKS
(1 - α)
record
don't record
record
don't record
in-degree
in-out
out-degree
in-degree
in-out
out-degree
DIR-R (PageRank)
DIR-UR (EigenFactor)
other
other
other
INDIR:DIR
UNDIR:DIR
OUTDIR-DIR (Count Links)

Smart Teleportation



Lambiotte, R. & Rosvall, M. (2012) Ranking and clustering of nodes in networks with smart teleportation

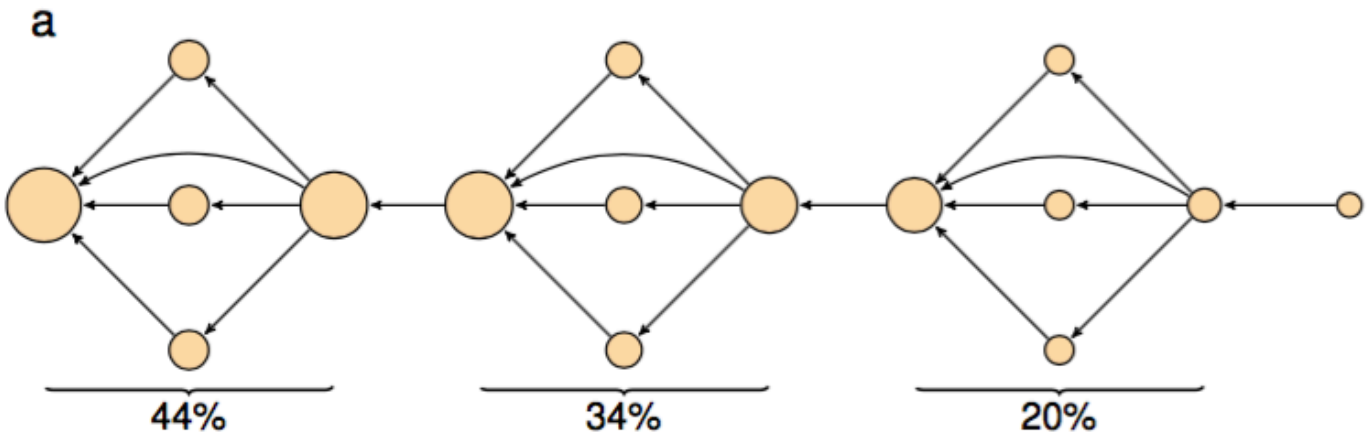
Smart Teleportation and Clustering



Lambiotte, R. & Rosvall, M. (2012) Ranking and clustering of nodes in networks with smart teleportation

Article-level Ranking and Mapping

DIR-R (**PageRank**)

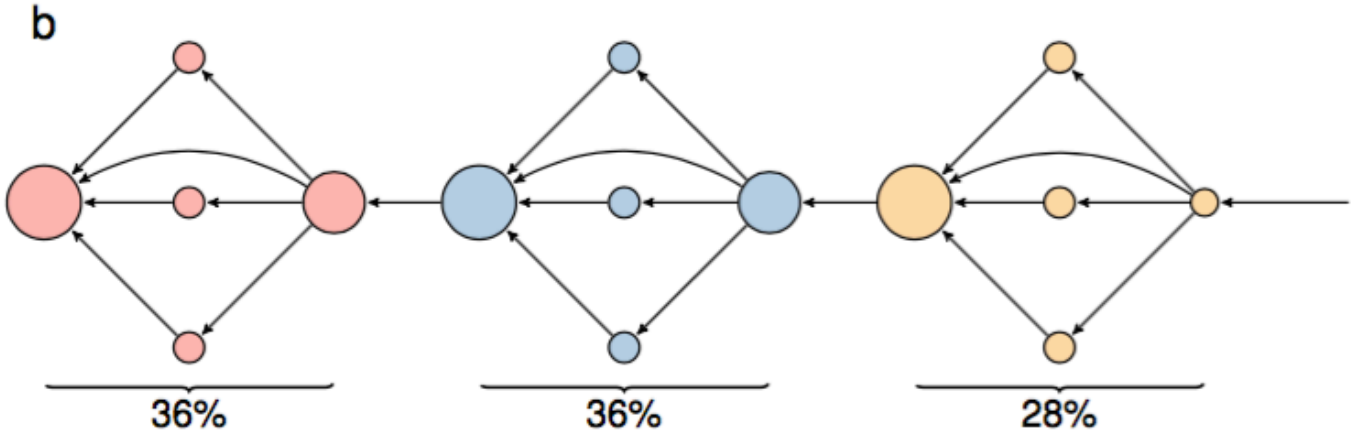


UNDIR:DIR

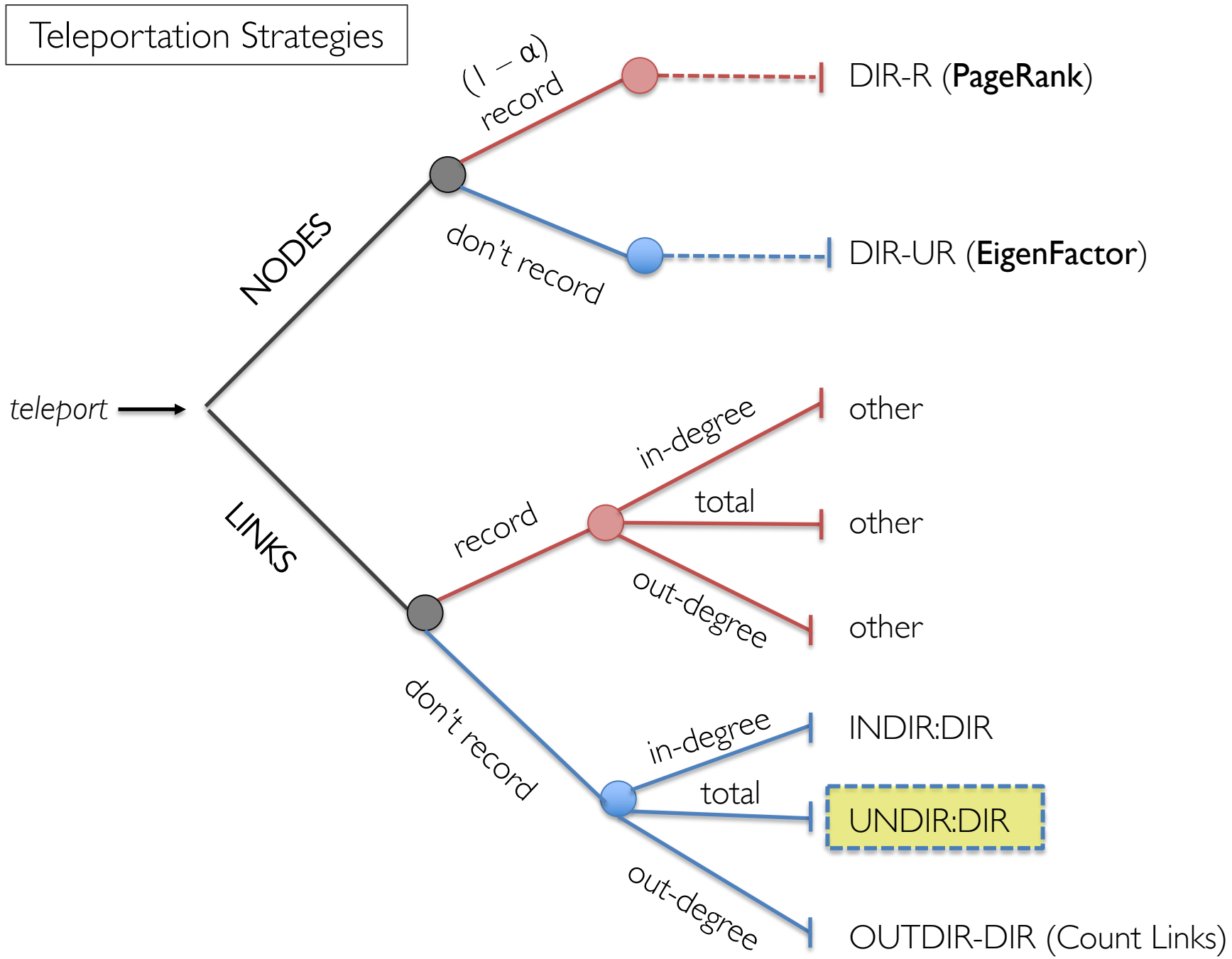
Smooths ranking

~

better clustering



West et al. (2016) Ranking and mapping article-level citation networks. *in prep.*


Teleportation Strategies
teleport
NODES
LINKS
(1 – α)
record
don't record
record
don't record
in-degree
total
out-degree
in-degree
total
out-degree
DIR-R (PageRank)
DIR-UR (EigenFactor)
other
other
other
INDIR:DIR
UNDIR:DIR
OUTDIR-DIR (Count Links)

Article-level Eigenfactor

$$w_i = \sum_j^n (Z_{ij} + Z_{ij}^T)$$

$$\mathbf{H}_{ij} = \frac{\mathbf{Z}_{ij}}{\mathbf{Z_i}}$$

$$\mathrm{ALEF} = n \frac{\mathbf{H}_{ij}^T . w_i}{\sum_i [\mathbf{H}_{ij}^T . w_i]_i}$$

Static Ranking of Scholarly Papers using Article-Level Eigenfactor (ALEF)

Ian Wesley-Smith
The Information School
University of Washington
Seattle, WA 98195 USA
iwsmith@uw.edu

Carl T. Bergstrom
Department of Biology
University of Washington
Seattle, WA 98195 USA
cbergst@uw.edu

Jevin D. West
The Information School
University of Washington
Seattle, WA 98195 USA
jevinw@uw.edu

ABSTRACT

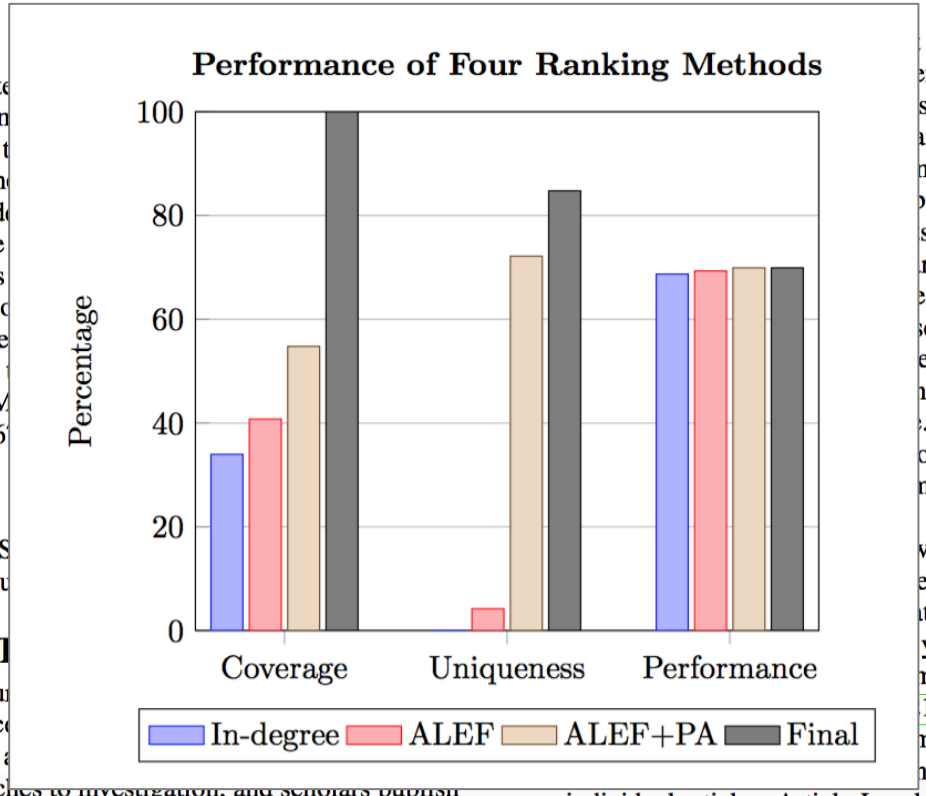
Microsoft Research hoste
on the Microsoft Acaden
rankings for the articles t
to be evaluated against th
Academic Graph provid
scholarly document, we
and used this contest as
Eigenfactor (ALEF), a nc
evaluate its performance
upon multiple facets of
(122M papers and 757M
contest was scored at 0.6

Keywords

Information Retrieval; S
works; Scholarly Commu

1. INTRODUCT

The scholarly literatu
and some estimates plac
scientific advances occur
a wealth of new approaches to investigation, and scholars publish
their work in a rapidly diversifying variety of venues. Despite all this,



material in the vast academic
entify documents that are both
sufficiently high importance to
ance is a problem of matching;
m of ranking. Once documents
by hyperlinks among them, the
s provides extensive contextual
nking. Google's PageRank al-
example. Before Google, most
search results without consider-
ey returned. Finding the most
n-curated web directories such
. Google showed us that latent
orld wide web lies the informa-
ns of determining relevance and

vn graph structure, with papers
e practice of scholarly citation.
te the process of scholarly dis-
years applying network-based
mine the importance of entities
1]. Although our methods were
portance of journals, we have
g methodology which operates
on individual articles: Article-Level Eigenfactor (ALEF) [27].
The 2016 WSDM Cup Challenge presented an excellent oppor-

WSDM CUP CHALLENGE

SIGN-UPS FOR THE WSDM CUP CHALLENGE ARE NOW CLOSED

The Graph

The Microsoft Academic Graph is a heterogeneous graph containing scientific publication records, citation relationships between publications, as well as authors, institutions, journal and conference "venues," and fields of study.

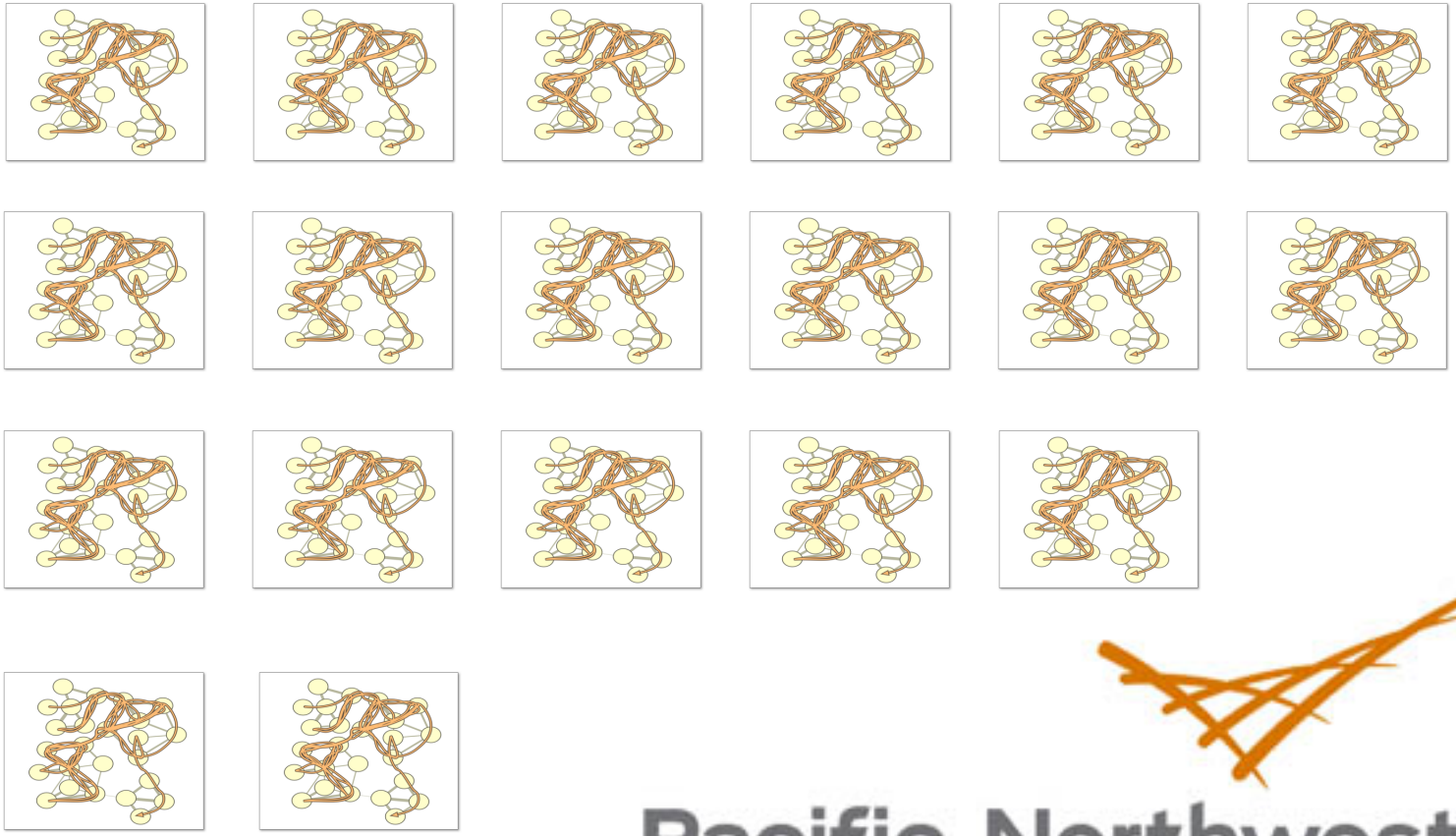
The Data

This data is available as a set of zipped text files stored in Microsoft Azure blob storage and available via HTTP. The file size (zipped) is ~30GB and may be downloaded here.

The Challenge

The goal of the Ranker Challenge is to assess the query-independent importance of scholarly articles, using data from the Microsoft Academic Graph.

Running Experiments

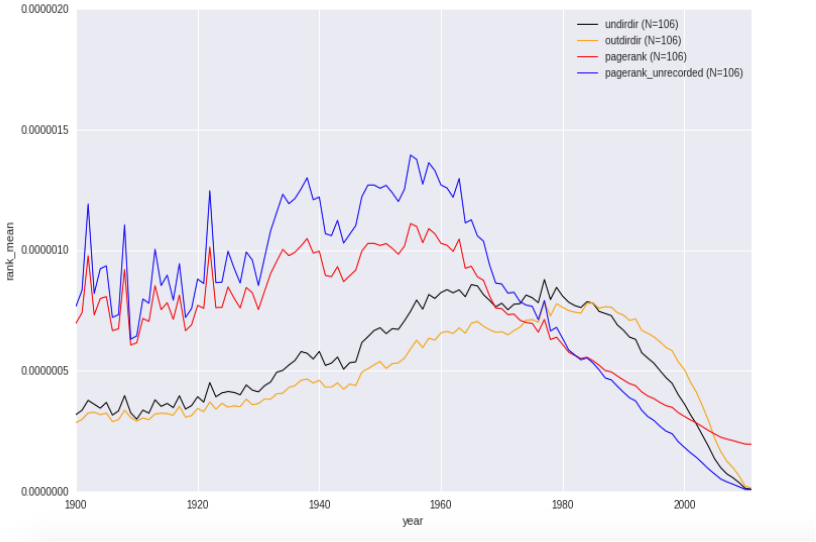


Pacific Northwest
NATIONAL LABORATORY

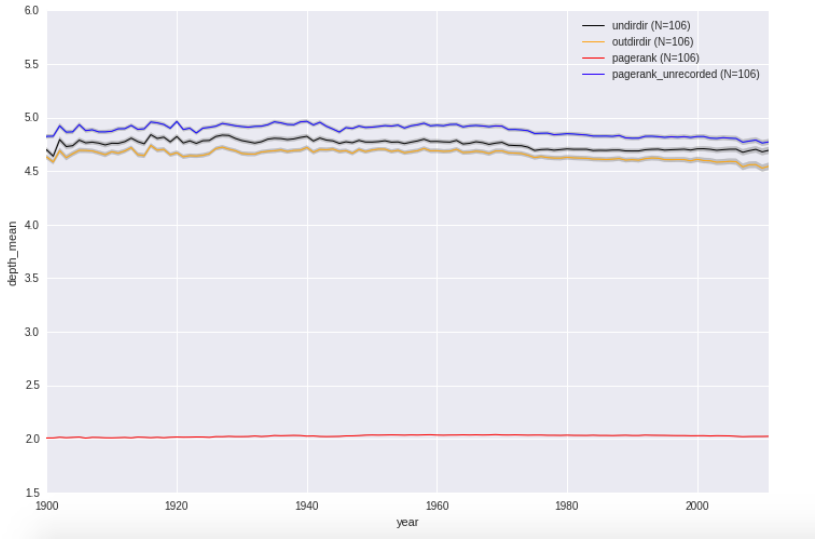
Clustering on time-directed networks

- Empirical exploration of hierarchical partitions with varying dynamics
- The effects of changing recorded teleportation ranking and clustering

Ranking Effects

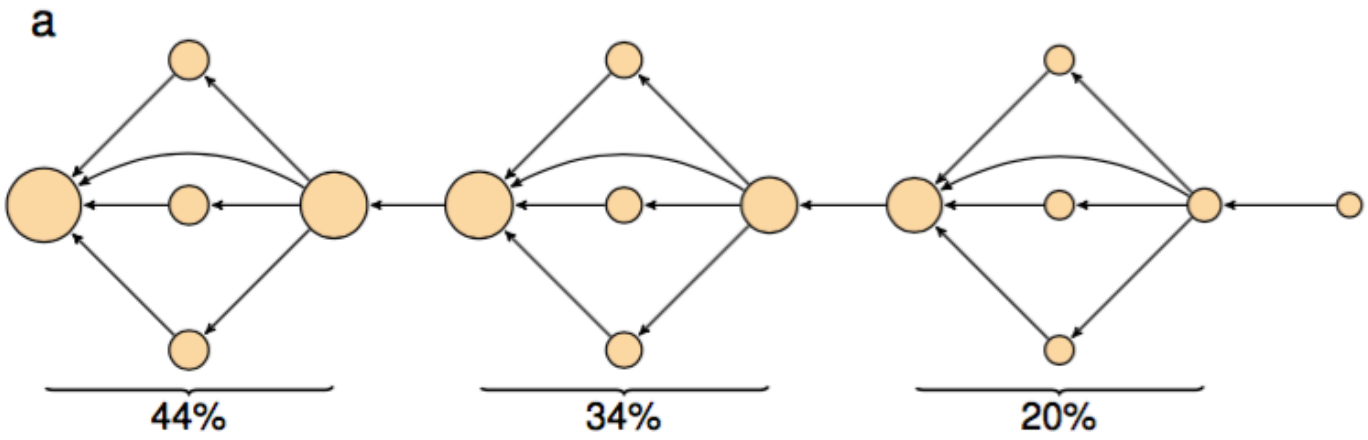


Clustering Effects



Article-level Ranking and Mapping

DIR-R (**PageRank**)

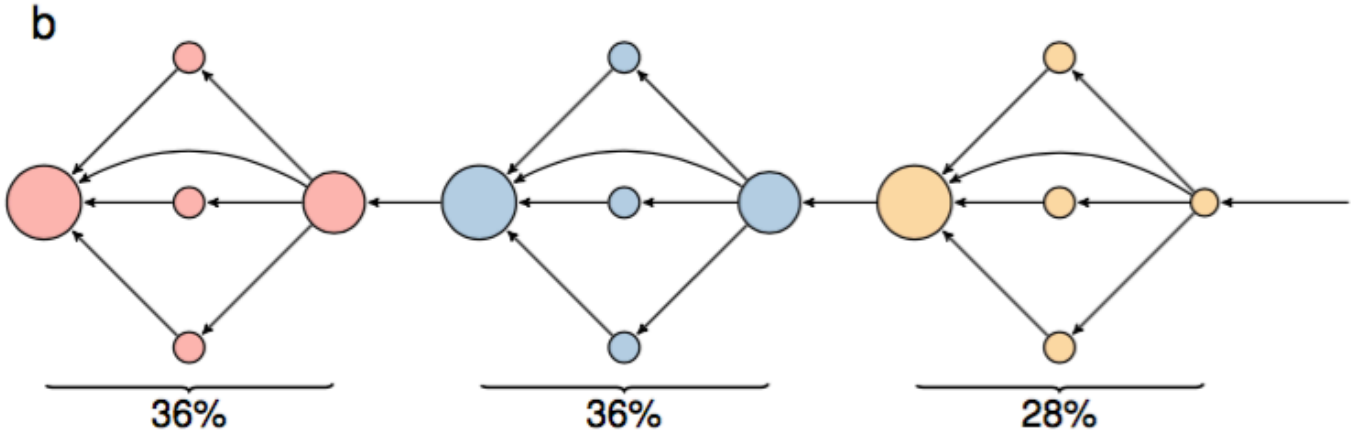


UNDIR:DIR

Smooths ranking

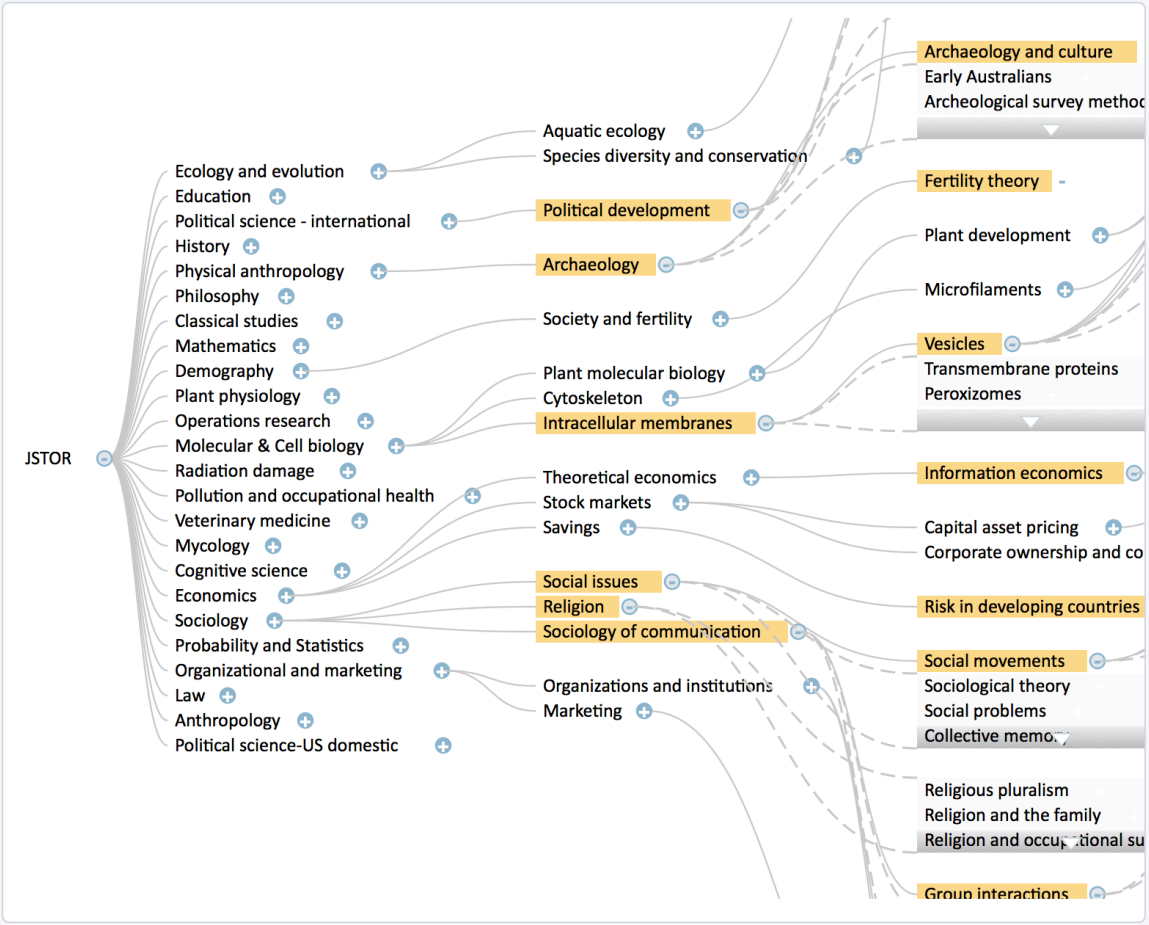
~

better clustering



West et al. (2016) Ranking and mapping article-level citation networks. *in prep.*

Revealing Hierarchical Structure



Find Papers

by **title** network

by **field**

by **author**

by **journal**

Active Queries: clear all

keyword: **network**

Top Papers Sort by Year (newest)

Using Siting Algorithms in the Design of Marine Reserve Networks
Heather Leslie - *Ecological Applications* (2003)

Mechanism of Filopodia Initiation by Reorganization of a Dendritic Network
Tatyana Svitkina - *The Journal of Cell Biology* (2003)

Network Structure and Knowledge Transfer: The Effects of Cohesion and Range
Ray Reagans - *Administrative Science Quarterly* (2003)

A General Model for Designing Networks of Marine Reserves
Enric Sala - *Science* (2002)

The Density of Social Networks and Fertility Decisions: Evidence from South Nyanza District, Kenya
Hans-Peter Kohler - *Demography* (2001)

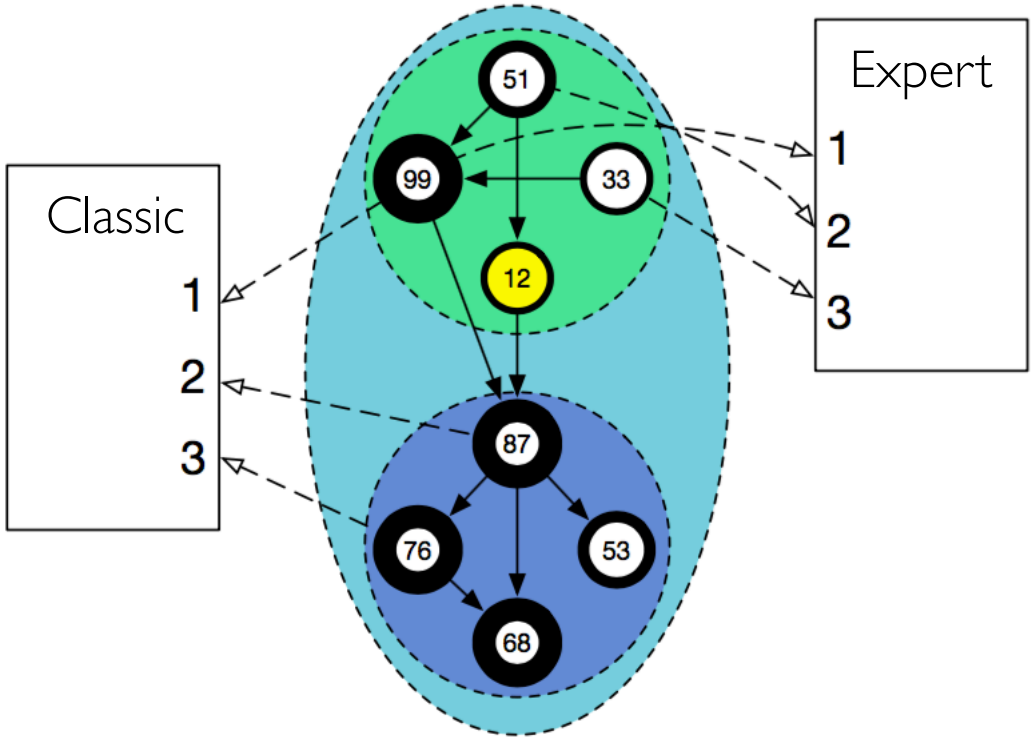
A New Dynamin-Like Protein, ADL6, Is Involved in Trafficking from the trans-Golgi Network to the Central Vacuole in Arabidopsis
Jing Bo Jin - *The Plant Cell* (2001)

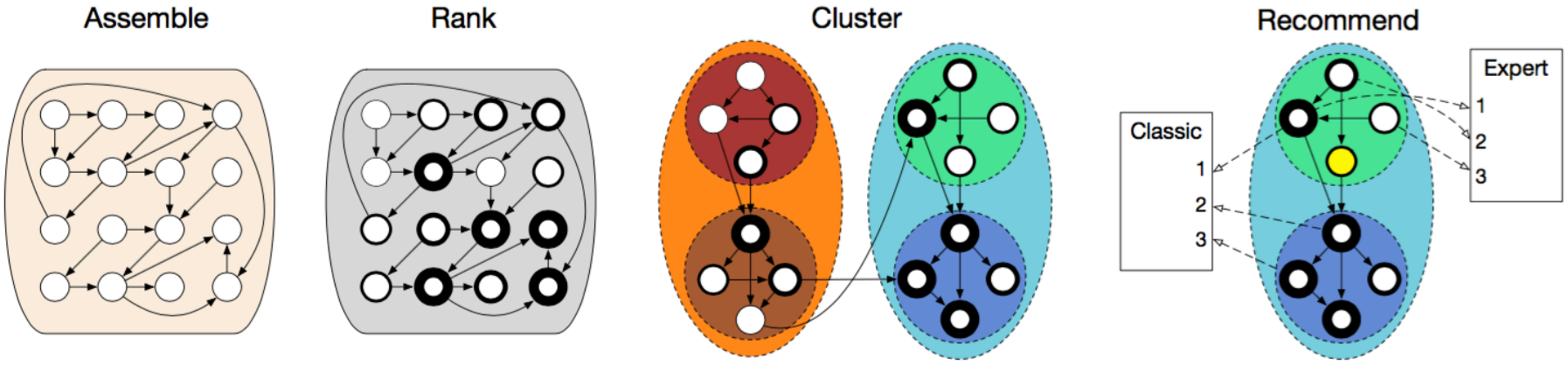
Comparing Sequenced Segments of the Tomato and Arabidopsis Genomes: Large-Scale Duplication Followed by Selective Gene Loss Creates a Network of Synteny
Hsin-Mei Ku - *Proceedings of the National Academy of Sciences of the United States of America* (2000)

A Noncooperative Model of Network Formation
Venkatesh Bala - *Econometrica* (2000)

JW

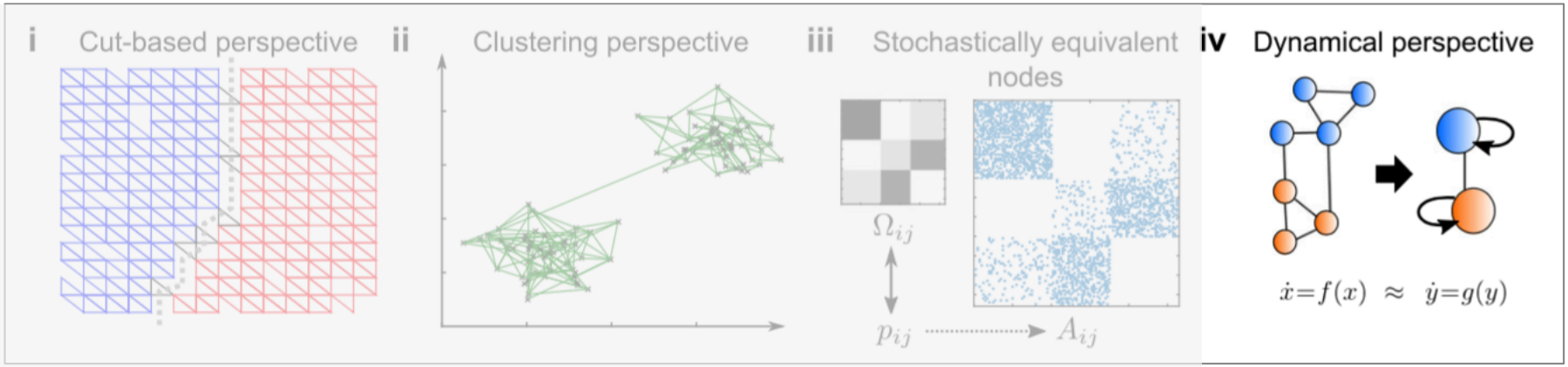
Recommend





West, Wesley-Smith, Bergstrom (2016) A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE, Transactions on Big Data* (in press)

Community Detection Perspectives



Circuit layout
Minimizing cuts
Load balancing
Eigenvectors
Spectral methods
Image segmentation

Data Clustering
Maximizing node density
unknown *k*, unbalanced
Conductance
Local, global
Modularity

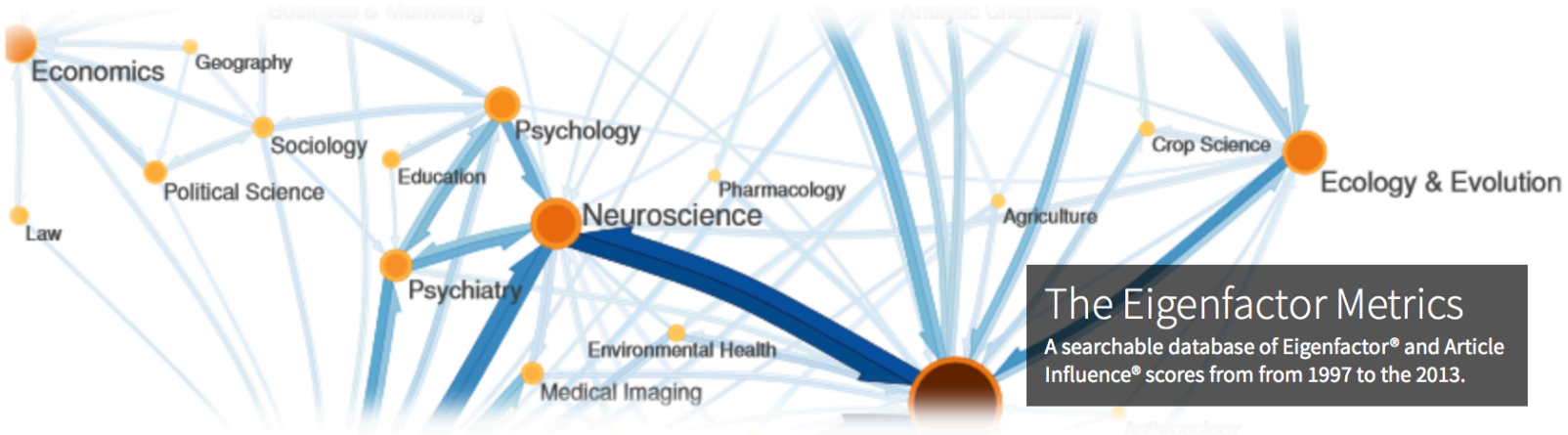
Social Networks
Connectivity Profiles
Stochastic equivalence
SBMs, LFR
p-values, hypothesis testing
Bipartite treatment
Predict missing links

System behavior, processes
Non-adjacency focused
Airline network
Markovian diffusion process
Undirected, Directed
InfoMap

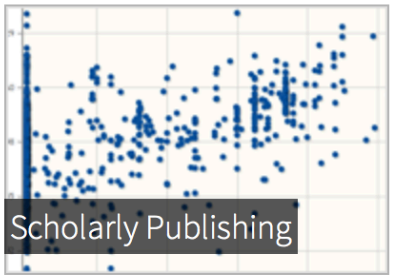
Schaub, M.T. et al. (2017) The many facets of community detection in complex networks. *Applied Network Science*

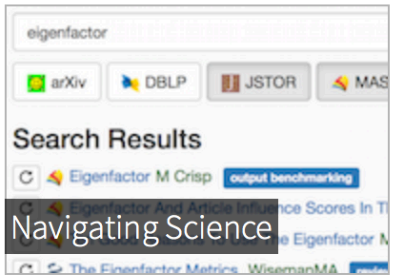
Summary

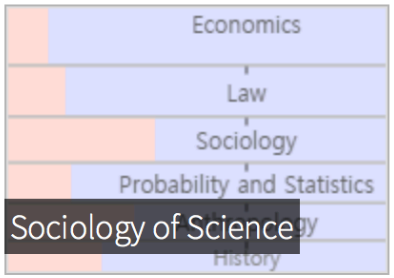
- Community detection – one size does not fit all
- Citation networks - dynamical perspective
- Memory - higher order dynamics
- Unrecordeded teleportation to links (*undirdir*) improves ranking and hierarchical clustering
- Next steps – building benchmarks and methods for evaluating the different rankings and hierarchical clusterings (refer to Jennifer Webster's talk tomorrow)



RESEARCH AREAS







NEWS

23 Nov.

JEVIN WEST ON MEGAJOURNALS IN THE *CHRONICLE OF HIGHER EDUCATION*

Jevin West discusses the rise of the megajournal and our open access cost effectiveness tool in the *Chronicle of Higher Education*.

---

23 Nov.

EIGENFACTOR TEAM PLACES SECOND IN MICROSOFT RESEARCH'S WSDM CUP

The WSDM Cup Challenge asked teams to use 30GB of data from the Microsoft Academic Graph to rank the

Acknowledgements

Carl Bergstrom, Department of Biology, University of Washington

Martin Rosvall, Department of Physics, Umea University

Seung-Hee Bae, Computer Science, Western Michigan

Jason Portenoy, Information School, University of Washington

Bill Howe, eScience, CSE, University of Washington

Jennifer Webster, Pacific Northwest National Laboratory

DataLab
Data Science and Analytics Lab

Jevin West

jevinw@uw.edu

jevinwest.org

@jevinwest