Alexander Troussov, Ph.D., IBM Dublin Software Lab

4th Russian Summer School in Information Retrieval, September 13-18, 2010, Voronezh

RuSSIR
Russian Summer School
in Information Retrieval

Graph-based Methods for Social Search

tutorial at RuSSIR 2010,
September 13-18, Voronezh, Russia

About AT

- IBM Ireland Center for Advanced Studies - Chief Scientist
- IBM LanguageWare group – the Architect
- National Geophysical Data Center, Boulder, CO, USA - Visiting scientist
  - Fuzzy logic based search engine for search in large databases when exact parameters of search are hard to define
- Observatoire de la Côte d'Azur, Nice, France – Visiting scientist
  - numerical simulation in stochastic physics.
- Institute of Physics of the Earth (Russian Academy of Sciences) and the International Institute for Earthquake Prediction Theory and Mathematical Geophysics, Moscow, Russia - Lead Researcher
  - R&D in geophysics and geoinformatics
- System programming at the Institute of Precise Mechanics, Moscow
- PhD in Mathematics from Lomonosov Moscow State University

- This tutorial is extension of my previous tutorials with Prof. Eugene Levner (HIT, Israel)

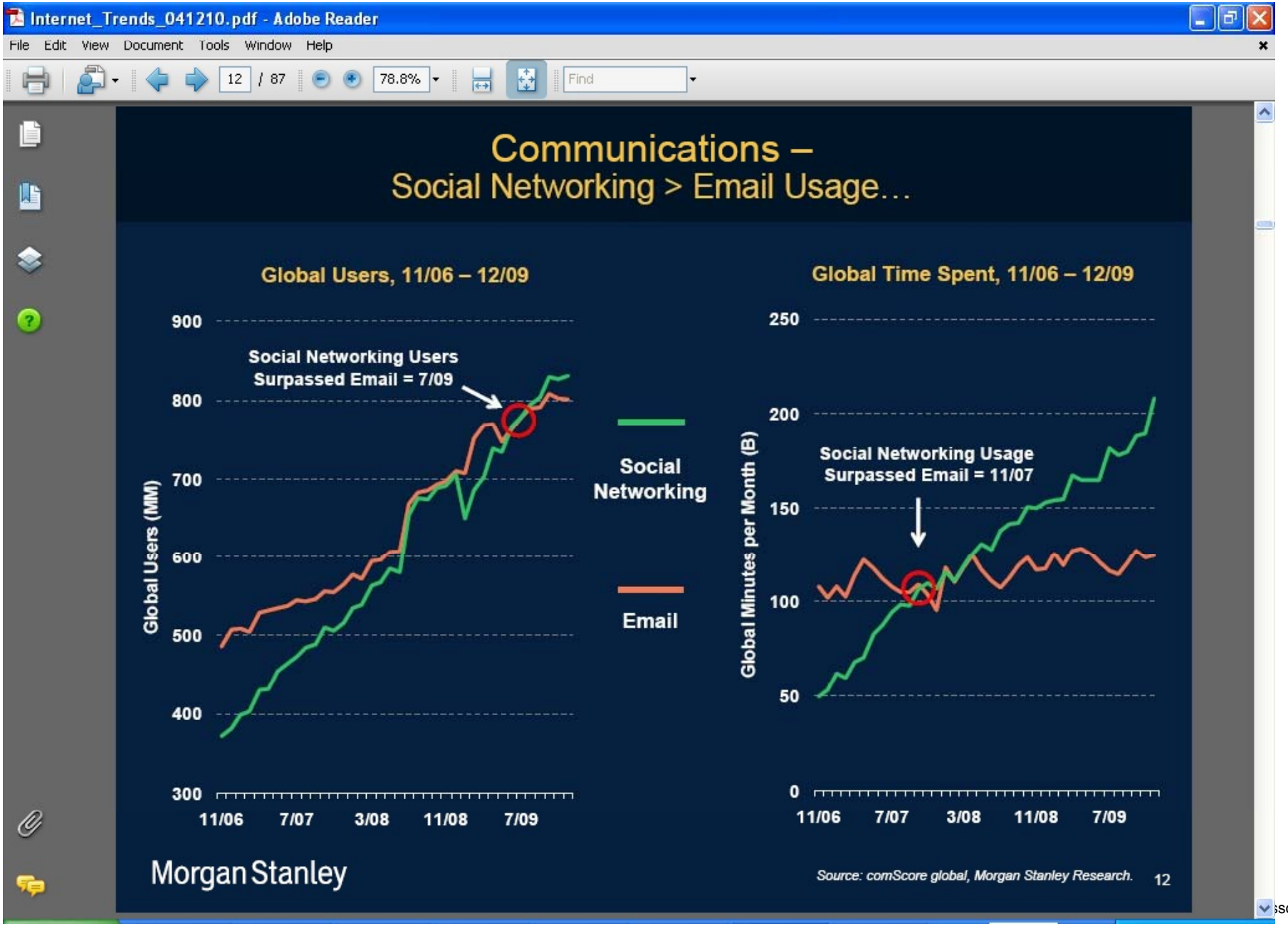
RuSSIR
Russian Summer School in Information Retrieval

Why Social Search?
Why graph-based Methods?

Why Social Search?

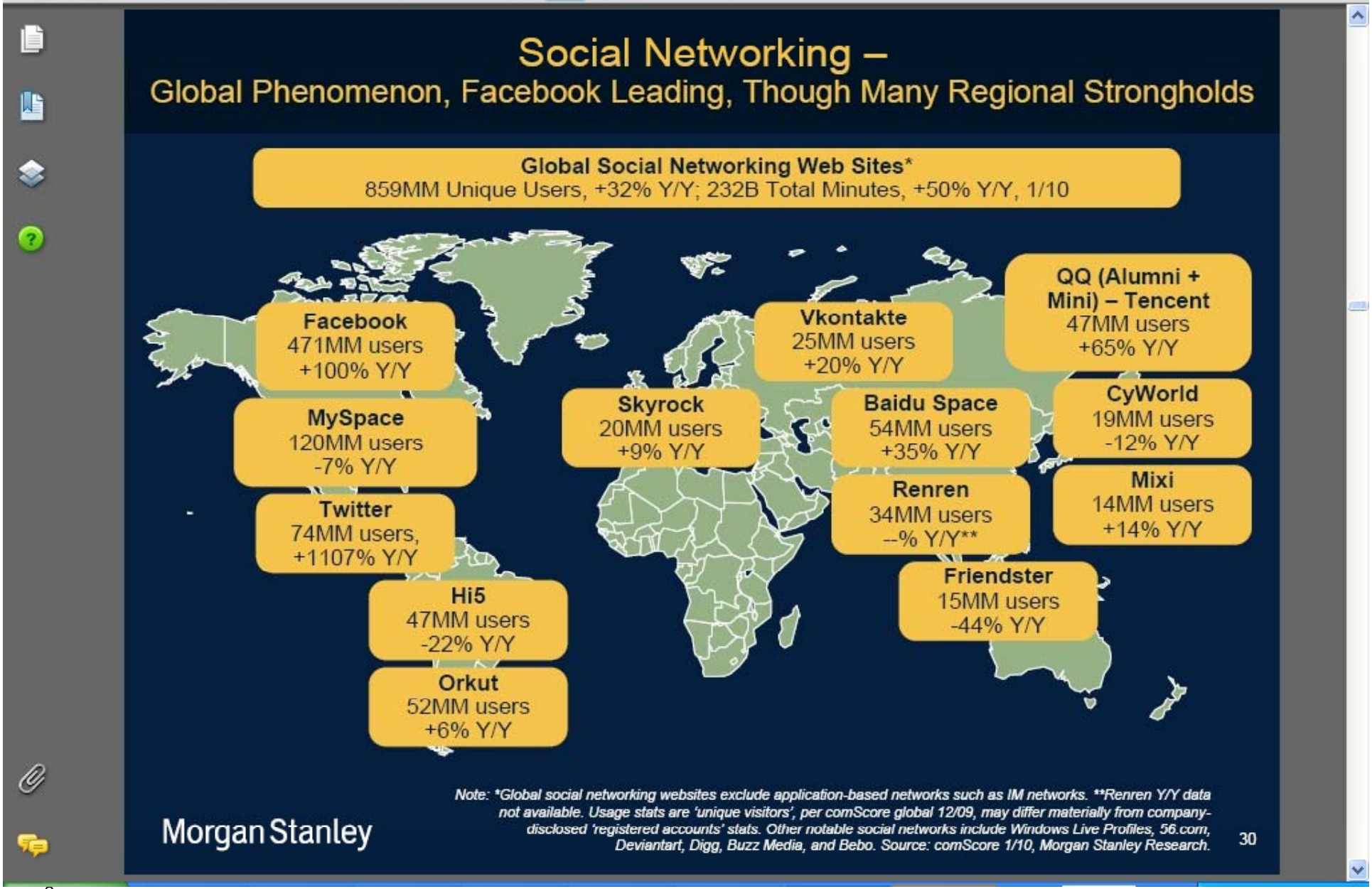
- Social and mobile (location, time) context improves search and information discovery experience
- Nielsen Search Report: March 22nd, 2010.
  - "a major influence of Facebook for changing trends in search engine market share. Facebook increase there traffic significantly in past few months“

    http://www.saiinternetmarketing.com/blog/nielsen-search-report-bing-increasing-search-market-share/


Internet_Trends_041210.pdf - Adobe Reader
File Edit View Document Tools Window Help
12 / 87
78.8%
Find
Communications –
Social Networking > Email Usage…
Global Users, 11/06 – 12/09
Global Users (MM)
900
800
700
600
500
400
300
Social Networking Users
Surpassed Email = 7/09
11/06
7/07
3/08
11/08
7/09
Social Networking
Email
Global Time Spent, 11/06 – 12/09
Global Minutes per Month (B)
250
200
150
100
50
0
Social Networking Usage
Surpassed Email = 11/07
11/06
7/07
3/08
11/08
7/09
Morgan Stanley
Source: comScore global, Morgan Stanley Research.
12

File Edit View Document Tools Window Help

30 / 87 78.8% Find



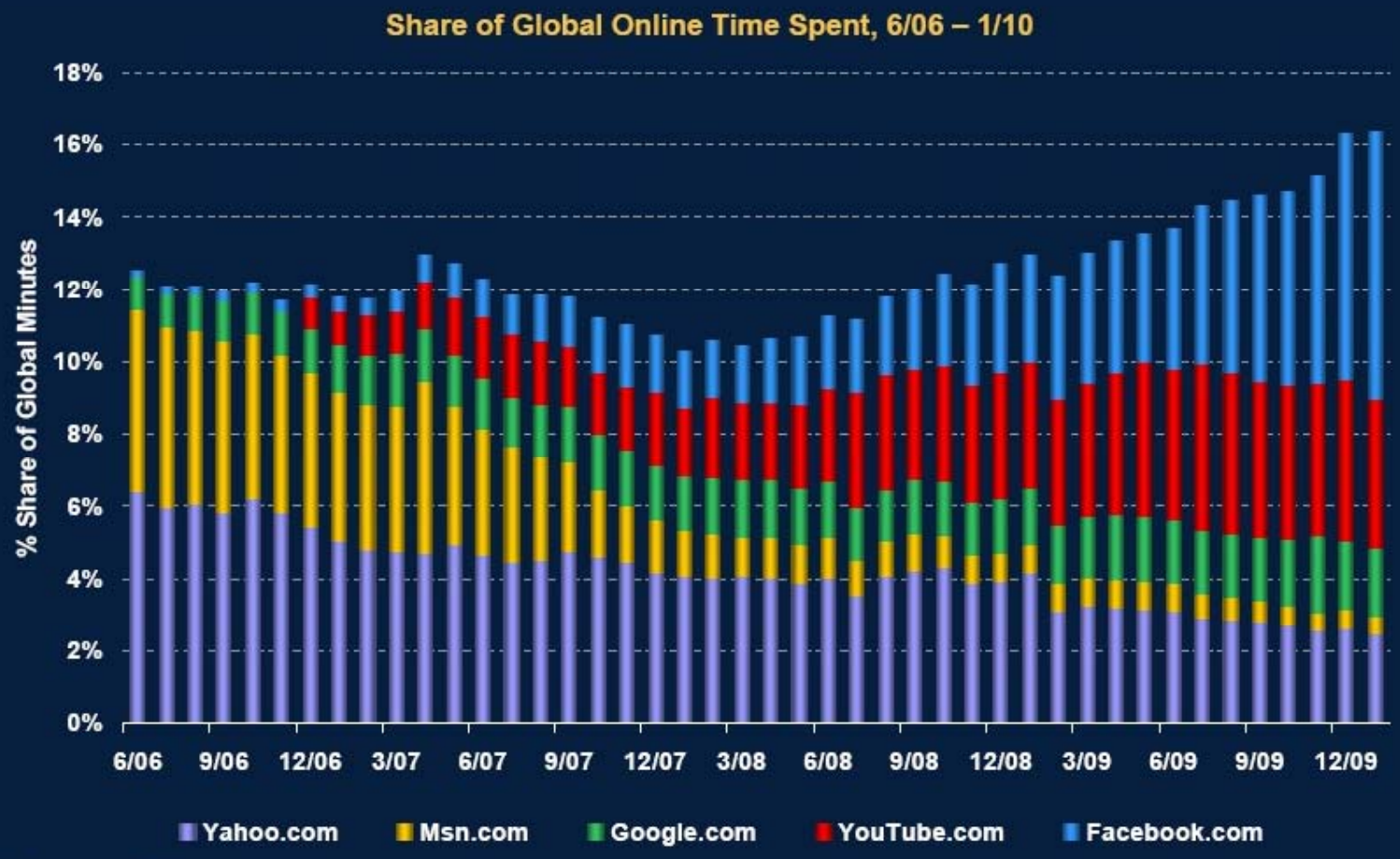
Source:

Internet_Trends_041210.pdf - Adobe Reader

File Edit View Document Tools Window Help

31 / 87 78.8% Find

Facebook (+ YouTube) = Largest Share Gainers of Global Online Usage Over Past 3+ Years



Morgan Stanley

Source: comScore global, 1/10.

Why Graph-based Methods?

- We have easily accessible network data on a large-scale
  - We live in an increasingly interconnected world of socio-technological systems, in which technological infrastructures composed of many layers are interoperating within a social context that drives their everyday use and development. Nowadays, most of the digital content is generated within public systems like Facebook, Delicious, Twitter, blog and wiki systems. These applications have transformed the Web from a mere document collection into a highly interconnected social space where documents are actively exchanged, filtered, organized, discussed and edited collaboratively.

    The emergence of the Social Web opens up unforeseen opportunities for observing social behaviour by tracing social interaction on the Web. In these socio-technological systems "everything is deeply intertwingled" using the term coined by the pioneer of the information technologies Ted Nelson: people are connected to other people and to "non-human agents" such as documents, datasets, analytic tools, tags and concepts. These networks become increasingly multidimensional providing rich context for network mining and understanding the role of particular nodes representing both people and digital content.

Source: Troussov et al. MITACS FP-Nets Workshop on Social Networks, Vancouver, Canada. 2010.

Representing knowledge

- There are a number of options:
  - As *objects*, using the well-accepted techniques of object-oriented analysis and design to capture a model
  - As *clauses* going back to the early days of AI and Lisp
  - As *XML*, using the industry-standard structured mark-up language
  - As *graphs*, making use of the things we know about graph theory
  - As some combination of these
- We are looking for
  - Extensibility
  - Easy of merge heterogeneous information
  - Ease of use
- **Graphs**
  - **We can use the nodes of a graph for facts, concepts, people, organizations, etc and the arcs as binary relationships between them**
    - **Arcs are typically called predicates or relationships in this view**
    - **The set of arcs intersecting a node tells us the information we know about the fact or entity**

Source: Simon Dobson, Declan O'Sullivan

Alexander Troussov, Ph.D., IBM Dublin Software Lab

4th Russian Summer School in Information Retrieval, September 13-18, 2010, Voronezh

RuSSIR
Russian Summer School
in Information Retrieval

Social Knowledge Management

The social context (the set of facts or circumstances that surround a situation or event) is hard to represent to our formal reason

Source: Prof. Noshir Contractor

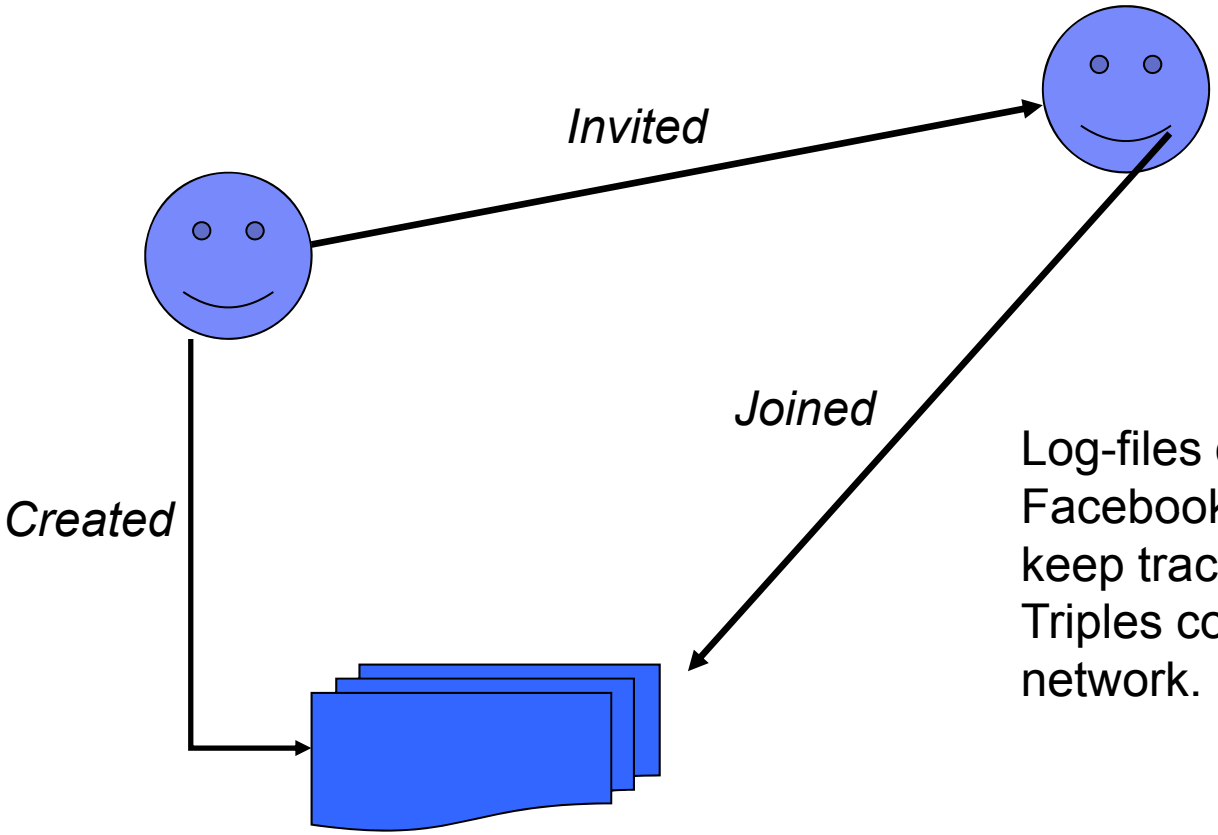
- The World Wide Web has changed almost every aspect of modern life. It touches us all. The Web's billions of pages, links and other resources comprise the largest information fabric in the history of humanity. Yet the Web is rarely approached as an object of scientific study. What processes have driven the Web's growth, and will they persist? How does large-scale structure emerge from a simple set of protocols? How does the Web work as a socio-technical system? What drives the viral uptake of certain Web phenomena? What might fragment the Web? Featuring some of the world's leading researchers on these areas this interdisciplinary meeting will discuss these and other issues as it presents the components of a Science of the Web.

  Web Science Trust

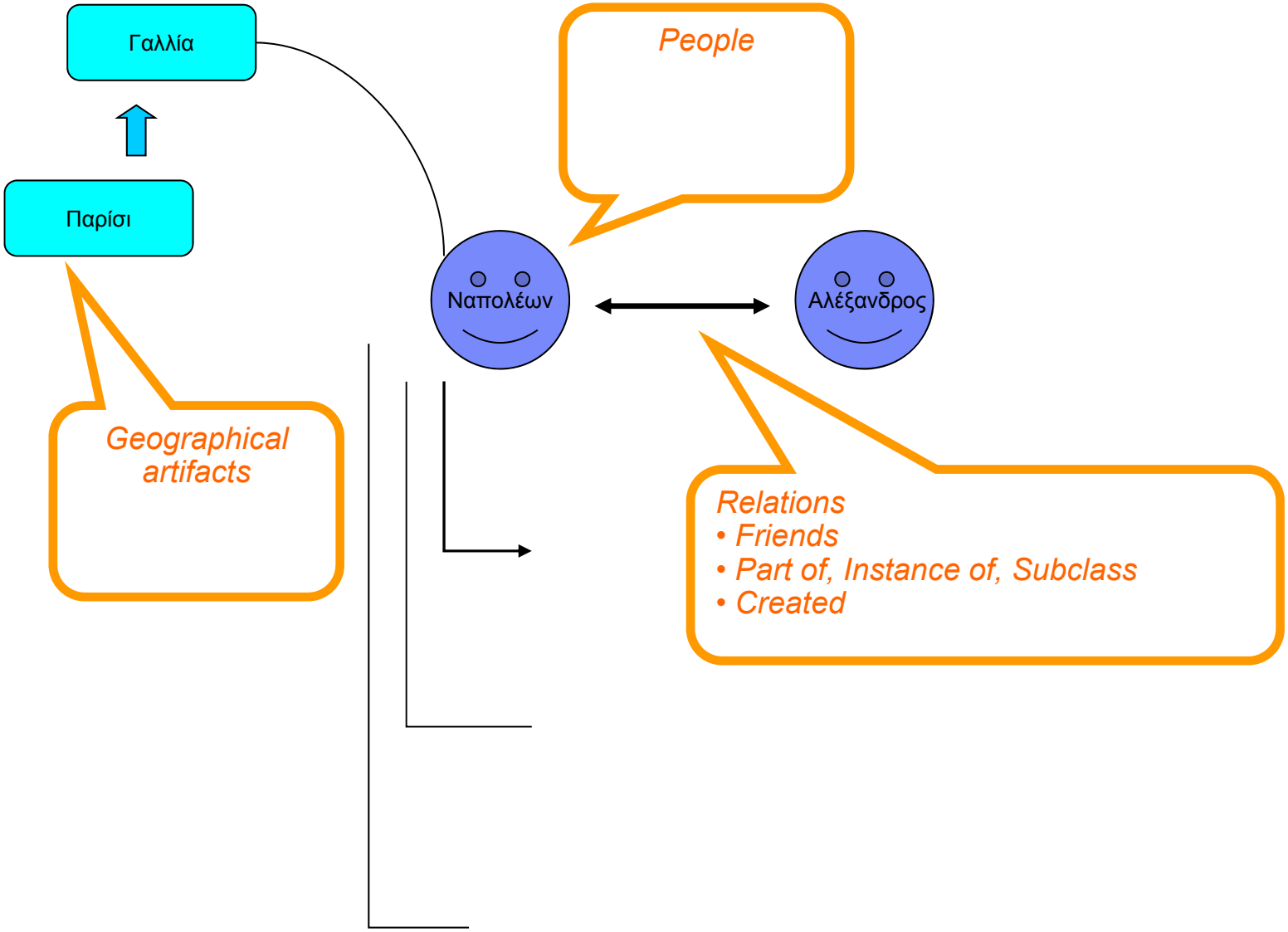
The social context: introduction

- We live in an increasingly interconnected world of techno-social systems, in which technological infrastructures composed of many layers are interoperating within a social context that drives their everyday use and development. Nowadays, most of the digital content is generated within public systems like Facebook, Delicious, Twitter, blog and wiki systems, and also enterprise environments such as Microsoft SharePoint, and IBM Lotus Connections. These applications have transformed the Web from a mere document collection into a highly interconnected social space where documents are actively exchanged, filtered, organized, discussed and edited collaboratively.
- The emergence of the Social Web opens up unforeseen opportunities for observing social behavior by tracing social interaction on the Web. In these socio-technological systems "everything is deeply intertwingled" using the term coined by the pioneer of the information technologies Ted Nelson: people are connected to other people and to "non-human agents" such as documents, datasets, analytic tools, tags and concepts. These networks become increasingly multidimensional providing rich context for network mining and understanding the role of particular nodes representing people and digital content.

How to model the social context?



Log-files of Techno-Social systems (like Facebook or IBM's Lotus Connections) keep track about *who did what*.
Triples could be aggregated into a network.



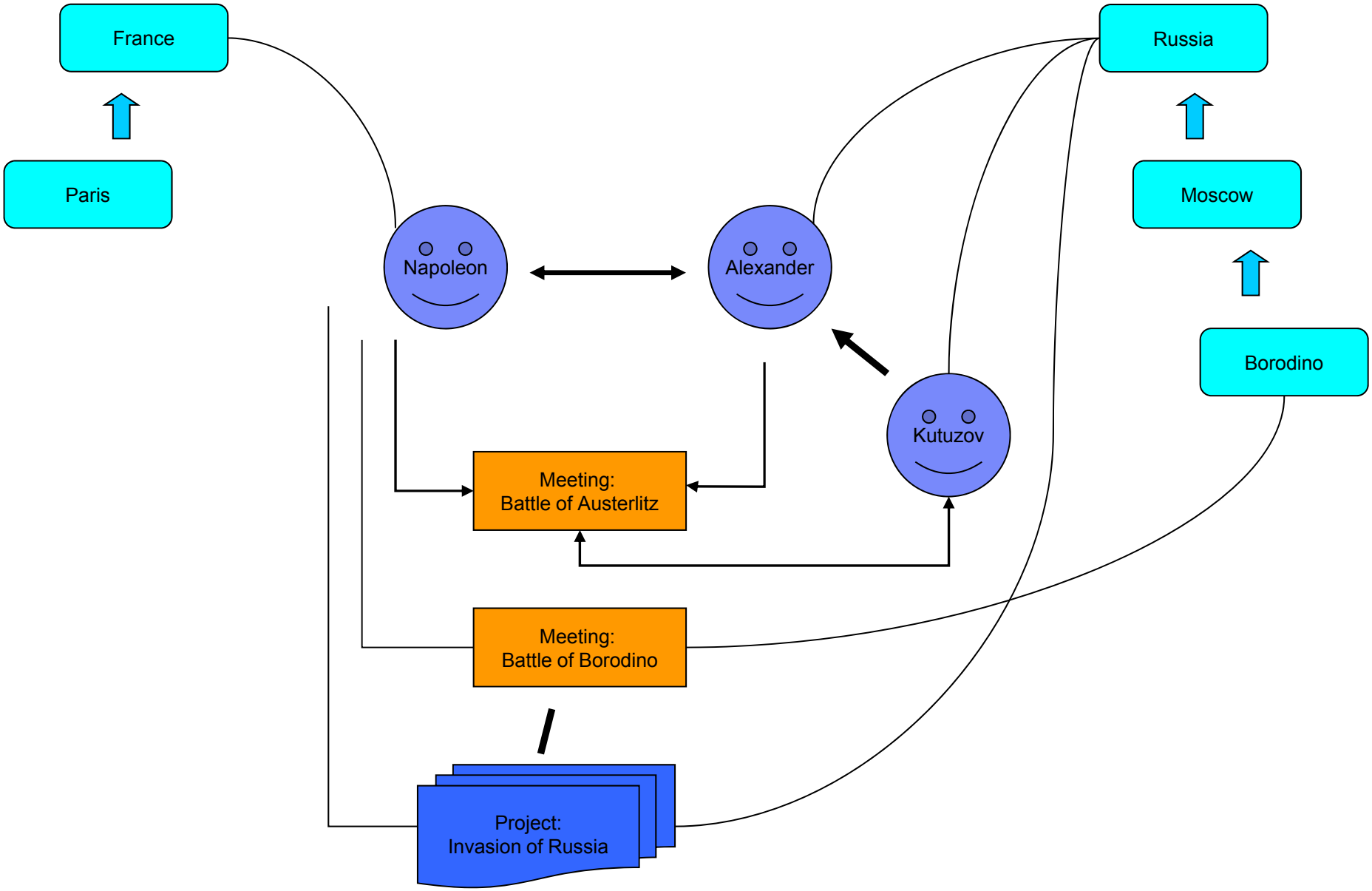
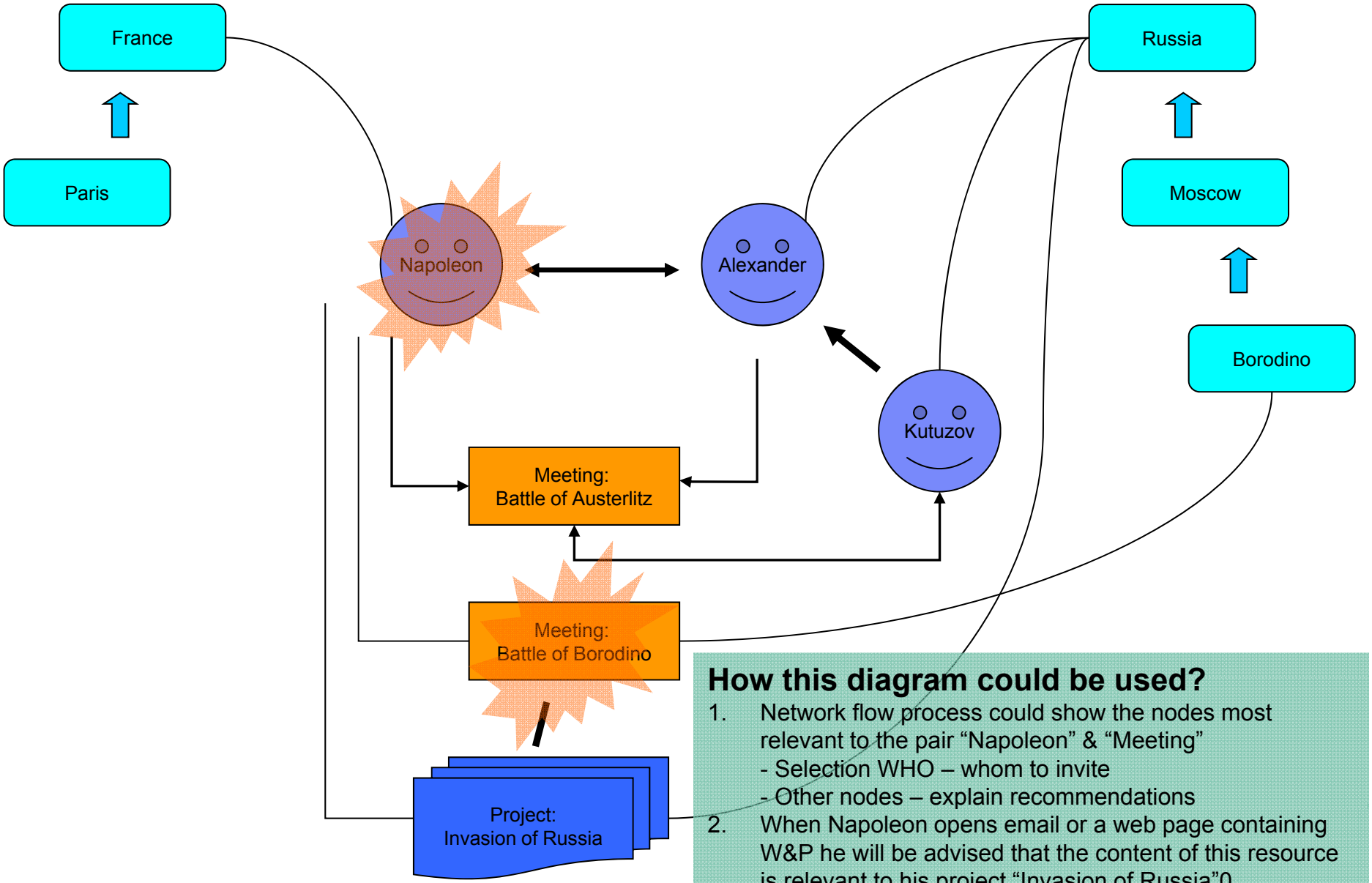


Diagram on the previous slide …

- What it represents ?
- How it can be used ?



How this diagram could be used?

1. Network flow process could show the nodes most relevant to the pair “Napoleon” & “Meeting”
   - Selection WHO – whom to invite
   - Other nodes – explain recommendations
2. When Napoleon opens email or a web page containing W&P he will be advised that the content of this resource is relevant to his project “Invasion of Russia”0

Diagram on the previous slide … What it represents?

- Data from Facebook, data from Napoleon's Lotus Notes calendar, structure of a Wiki, network of collocations or relations between the entities in W&P, …
  - The proliferation of Web 2.0 and Enterprise 2.0 technologies has lead to the emergence of massive networks connecting people and various digital artifacts. These networks can be treated as a "weak" knowledge, which nevertheless might be used recommendations and even for such traditional applications as knowledge-based text processing
- Or instantiation of an ontology related to W&P by Leo Tolstoy
  - In which case we would probably know that Napoleon is emperor of France, Paris is the capital (not instantiation of a subclass) of France, etc.
- Ontology provides conceptualization, allow inferencing, but these advantages per se are useless without tedious manual work to encode the rules how to use this additional knowledge. While the knowledge encoded in the topology of the multidimensional network is ready to use provided that methods are tolerant to errors and inconsistencies in data - i.e. the methods are methods of "soft mathematic" – fuzzy inferencing, soft clustering, …

Social Context = (Social) Knowledge ?

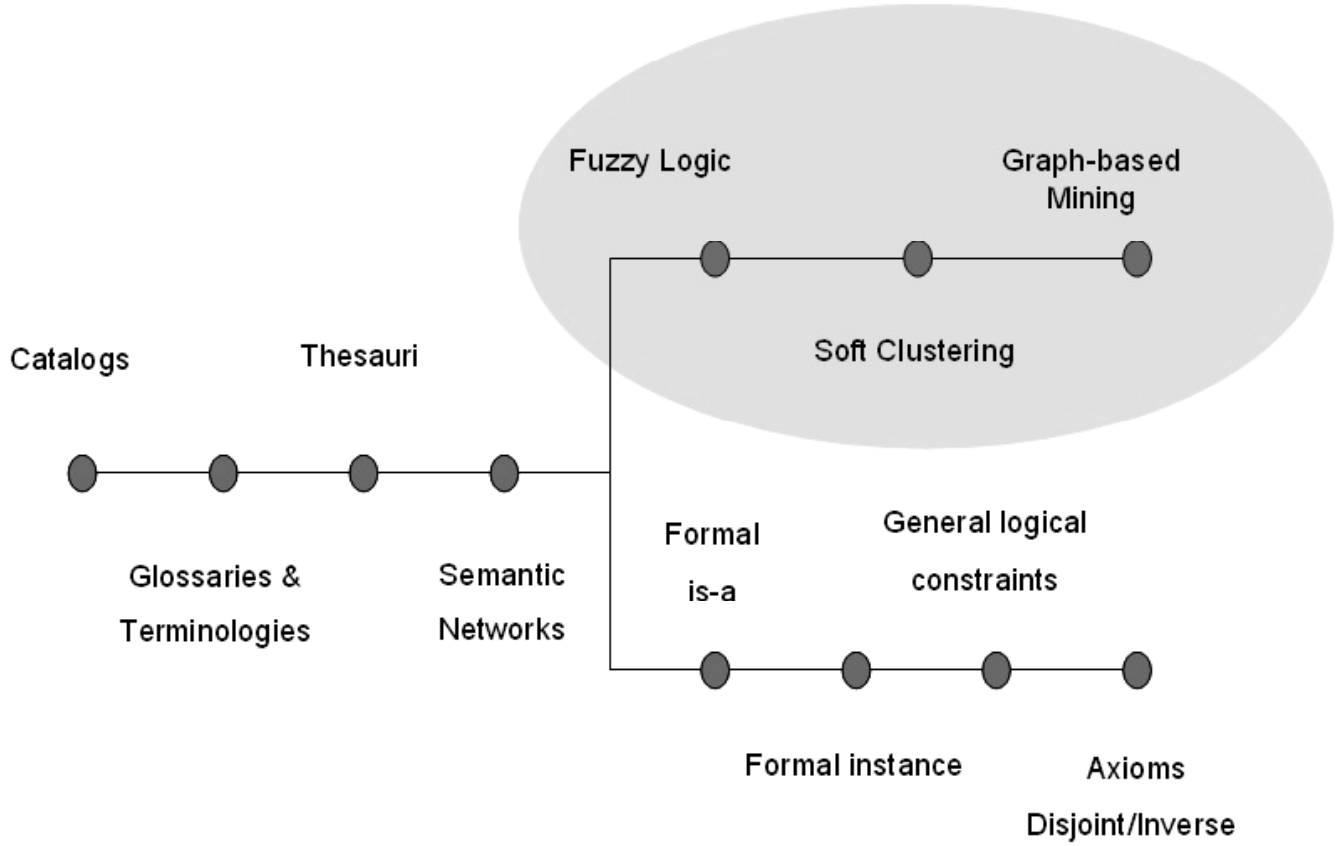
A New Mathematical Model of Horse Racing

- *Assume, without the loss of generality, that each horse in the horse racing is modelled by a wooden ball of radius R_i.*

= a ball ? ☺

Representing social context as a knowledge allows us to benefit from the past experience of knowledge based applications

For instance, the social context modeled as a network is not much different from semantic networks which are formed from concepts represented in ontologies. And it is possible to use such networks for knowledge based text processing. Representing social context as knowledge allows us to draw experience from such mature R&D area as knowledge-based text processing



Knowledge Management: To Codify or to Collaborate?

- Scientists talk about
  traditional dichotomy between codification and collaboration
  - – Codification: objectivist view of knowledge management as codification of knowledge in databases
  - – Collaboration: limited behaviourist view as communication and collaboration tools in support of social practices. …

  From the point of view of the traditional dichotomy between codification and collaboration approaches to knowledge management, the social context could be considered as bottom-up created social knowledge.

To Codify or to Collaborate?

- Codification vs. use (business practices)
  - Traditionally,
    - Business looks for local benefit
    - The State provides infrastructures for common good
- Codification – is an infrastructure which does not provide local benefits
  - WordNet provides benefits for its user, not necessarily to its creators
  - Private companies have not enough incentives to invest in codification
    - (at least if they don't have monopoly)
- Codification can be very efficient
  - If done by a monopolist. State has monopoly for "common good"
    - state sponsored research -> codification –> business use -> common good

To Codify or to Collaborate? (Cont.)

- Conclusions:
  - Codify – Prescribe approach
    - Is a long term investment
  - Collaborate - Describe approach
    - Has immediate gain (in the true spirit of Web 2.0)
  - Probably, business will not be very much interested “to codify”

Dichotomy? Two types of knowledge ?

- Use case of knowledge based text processing
  - where we can find “the right” knowledge for recommender systems, for better search, for text processing?
  - Right for the domain
    - Knowledge about Symantec Client Firewall is not suitable to process W&P
  - Right for computers

Ontology? Data from calendar, or from Facebook?

- The proliferation of Web 2.0 and Enterprise 2.0 technologies has lead to the emergence of massive networks connecting people and various digital artifacts. These networks can be treated as a “weak” knowledge, which nevertheless might be used for knowledge based text processing
  - Bad news – we don’t have enough “proper” knowledge
  - Good news – the amount of “weak” knowledge is enormous

KM Practitioners: KM goes social

- Old fashion KM (KM 1)
  - Technocentric; driven by IT; very top-down
    - captures unstructured info and stores it in DB
  - Failed to leave to its expectations
- Social tools
  - bottom-up; decentralized; less monolithic
- KM (2.0) goes Social
  - KM 2.0 ?
    - "I don't know if it's going to be dubbed that“
      social tools / knowledge sharing tools - alternative labels

David Gurteen on KM 2.0 (2007)

Has KM died, and resurrected as social computing?

- Has KM died, and resurrected as social computing?
  - This was alluded to back in 2005 by Martin Dugage, who said businesses were looking at KM as, “processes, engineering and mechanics”, rather they should be looking at it as, “practices, creativity, and social networks.”

Social knowledge - Wisdom of crowd

- Social knowledge has recently attained recognition (and value) under the heading of Surowiecki's 'wisdom of crowds'.
  - But it is worth noting that many of Surowiecki's examples are cases where individual guesses "aggregated and then averaged." While Surowiecki stresses (correctly) the autonomy of those guesses, he does not so stress the equally important fact that those guesses are not independent events - they are connected, in some key way, to each other (for example, the people guessing the temperature of a room have also the property of being in the same room; those estimating the weight of objects all see the same objects, and in the same way).

James Michael Surowiecki is an American journalist. In 2004, he published The Wisdom of Crowds, in which he argued that in some circumstances, large groups exhibit more intelligence than smaller, more elite groups, and that collective intelligence shapes business, economies, societies and nations.

Wikipedia

Social knowledge - Wisdom of crowd (Cont.)

- **Social knowledge is not merely the aggregation and averaging of individual knowledge** (as if there could be such a thing - consider how in guessing weights we use a medium, which in electing leaders we use a mode). That is why such aggregation is not necessarily reliable - an aggregation that is considered independently of the connections between entities is like a count that is considered independently of the membership of a set. Consider, for example, counting sheep without worrying about whether what is being counted is a sheep. It can work sometimes - in sheep-filled rooms, for example. But more often, it will mislead." (a result of the connections between the individual members of society, resident in no single one of them, but rather a property of the society working as a whole)

Social Knowledge

- Stephen Downes
  - **"Social knowledge is to a society what personal knowledge is to a person.** It is a result of the connections between the individual members of society, resident in no single one of them, but rather a property of the society working as a whole. Numerous instances of such connections occur; where certain of those connections become salient, and are frequently activated through use, they are recognized as forming a distinct entity, producing a distinct type of knowledge.
    (From P2P Foundation http://p2pfoundation.net/Social_Knowledge)
- Etienne Wenger
  - **Traditional knowledge management approaches attempt to capture existing knowledge within formal systems, such as databases. Yet systematically addressing the kind of dynamic "knowing" that makes a difference in practice requires the participation of people who are fully engaged in the process of creating, refining, communicating, and using knowledge.**

Dimensions of techno-social systems

- The power of social network analysis stems from its difference from traditional social scientific studies, which assume that it is the attributes of individual actors—whether they are friendly or unfriendly, smart or dumb, etc.—that matter.
- Social network analysis produces an alternate view, where the attributes of individuals are less important than their relationships and ties with other actors within the network.
  - The power of social network analysis stems from its difference from traditional social scientific studies, which assume that it is the attributes of individual actors—whether they are friendly or unfriendly, smart or dumb, etc.—that matter.
  - This approach has turned out to be useful for explaining many real-world phenomena, but leaves less room for individual agency, the ability for individuals to influence their success, because so much of it rests within the structure of their network.
  - However, this defect of traditional SNA can be overcome, if we consider multidimensional networks

Socio-semantic networks

- *Everything is deeply intertwingled. In an important sense there are no "subjects" at all; there is only all knowledge, since the cross-connections among the myriad topics of this world simply cannot be divided up neatly.*
  Theodor Holm Nelson, the pioneer of information technology.

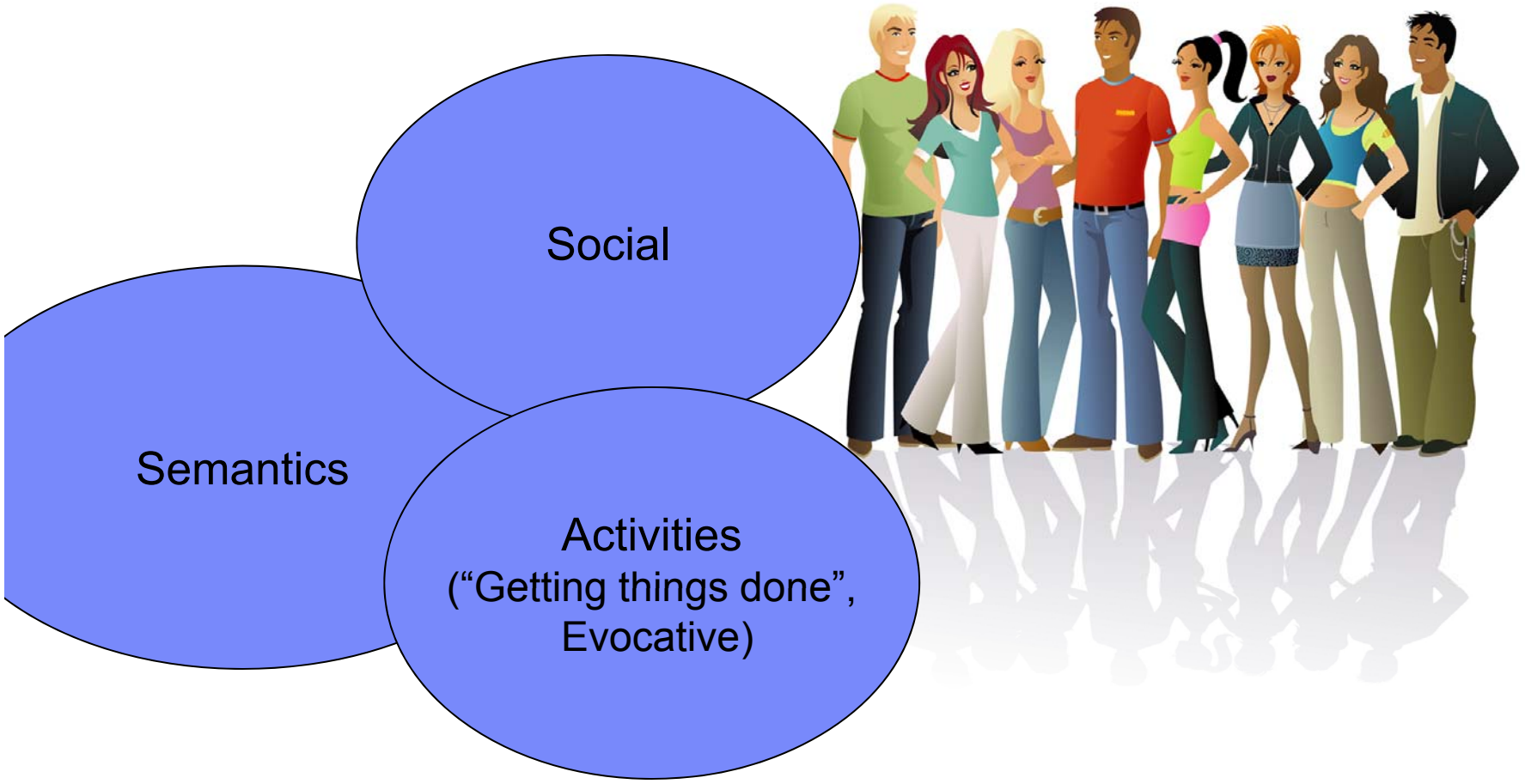
- and nowadays many of this links are explicit in computer mediated communication networks (the web, semantic web, desktop data, intranets, …), which can be studied using interdisciplinary methods

- We agree with J.R. Firth *"You shall know the word by the company it keeps"*, then nowadays we can *know the* ***text, person, project, "idea" etc*** *by the company they keep in the unified socio-semantic space.*

Socio-technical systems

- Researchers “complain”:
  - Emails are used for functions (like task management) it was not designed for
  - People use wrong tags in folksonomies
    - *“Many Zotero users use tags like 'need to read', 'skimmed', 'need to scan', etc. to organize their research library”*
- Before mining of socio-semantic networks,
  - let us think what people might bring to these networks.
  - let us first think about “dimensions” of computer mediated networks where people are involved
    - Web and other computer mediated networks are evolving extremely fast. Getting dimensions right will help us to see trends

Three dimensions of human life



Troussov and Kohlhase (2008)

Three pillars and one fastening ring of networks

- Semantics
  - After all, we call ourselves "***homo sapiens***" meaning "Man the Wise"
  - Semantics
    - Semantic web technologies
    - Traditional AI, including Natural Language Understanding
- Social
  - We are social beings as well as individuals / "To live in a society and be free from it is impossible" / Sometimes we want to be "***homo ludens***" (the "playing man")
- Activities management
  - "***Homo faber***" (Latin for "Man the Smith" or "Man the Maker")
  - This is about evocation, "getting things done", action management, etc
- Human Computer Interaction (HCI)
  - The proliferation of Web 2.0 has lead to the emergence of massive *networks connecting people and various digital artefacts*. The efficiency of human navigation in such networks depends on the availability of suitable user interfaces powered by an "intelligent" backend which provides guidance and recommendations. As not all of us all the time are knowledge workers, we'll probably need to move beyond the desktop metaphor.

Troussov and Konnase (2008)

Computer mediated networks are based on

Human Computer Interaction Troussov and Kohlhase (2008)

Conclusions:
Social Knowledge Management and Graph-based Methods

- Social-knowledge management
  - **Social knowledge management =**
    = social {knowledge management} + {social knowledge} management
  - Extensive use of “weak” knowledge

Conclusions (Cont.)

- Data are vied as a multidimensional network
  - To preserve context
- Mining is done by graph-based algorithms
  - Which perform well on multidimensional networks
  - and are very well aware of dimensions of computer mediated networks where people are Involved

Key element of the success

- **Avenues to deep socio-semantic analytics and the possibility of high-quality functionalities for socio-technical systems** (like recommending people to invite into your social network) hinge on the availability of engines which are able
  - to provide **hidden knowledge discovery** (like discovering a new relation in a network - that based on the strength of multiple connectivity between the nodes of a social network one can conclude that **Dr. Jekyll is related to Mr. Hide**),
  - and provide **ad hoc generalisation across dimensions**. For instance, the ability to detect that a particular person might serve as an representative of a community or as an expert on a particular topic (the example of such generalisation is the expression frequently attributed to Louis XIV "L'e'tat s'est moi (I'm the State).")

Alexander Troussov, Ph.D., IBM Dublin Software Lab

4th Russian Summer School in Information Retrieval, September 13-18, 2010, Voronezh

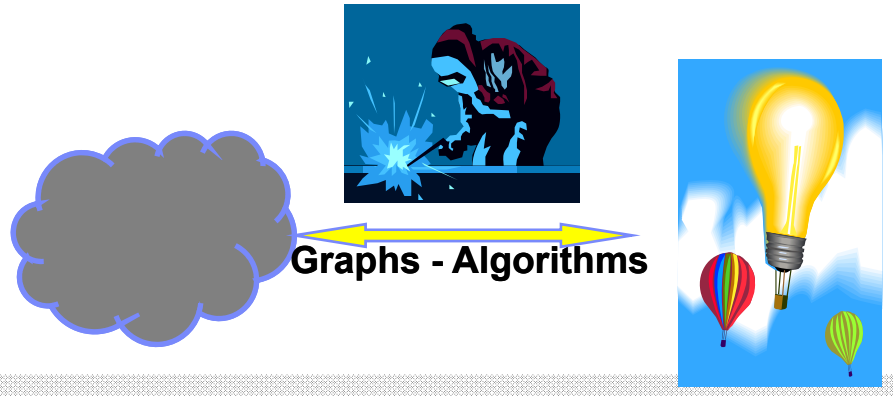
RuSSIR
Russian Summer School
in Information Retrieval

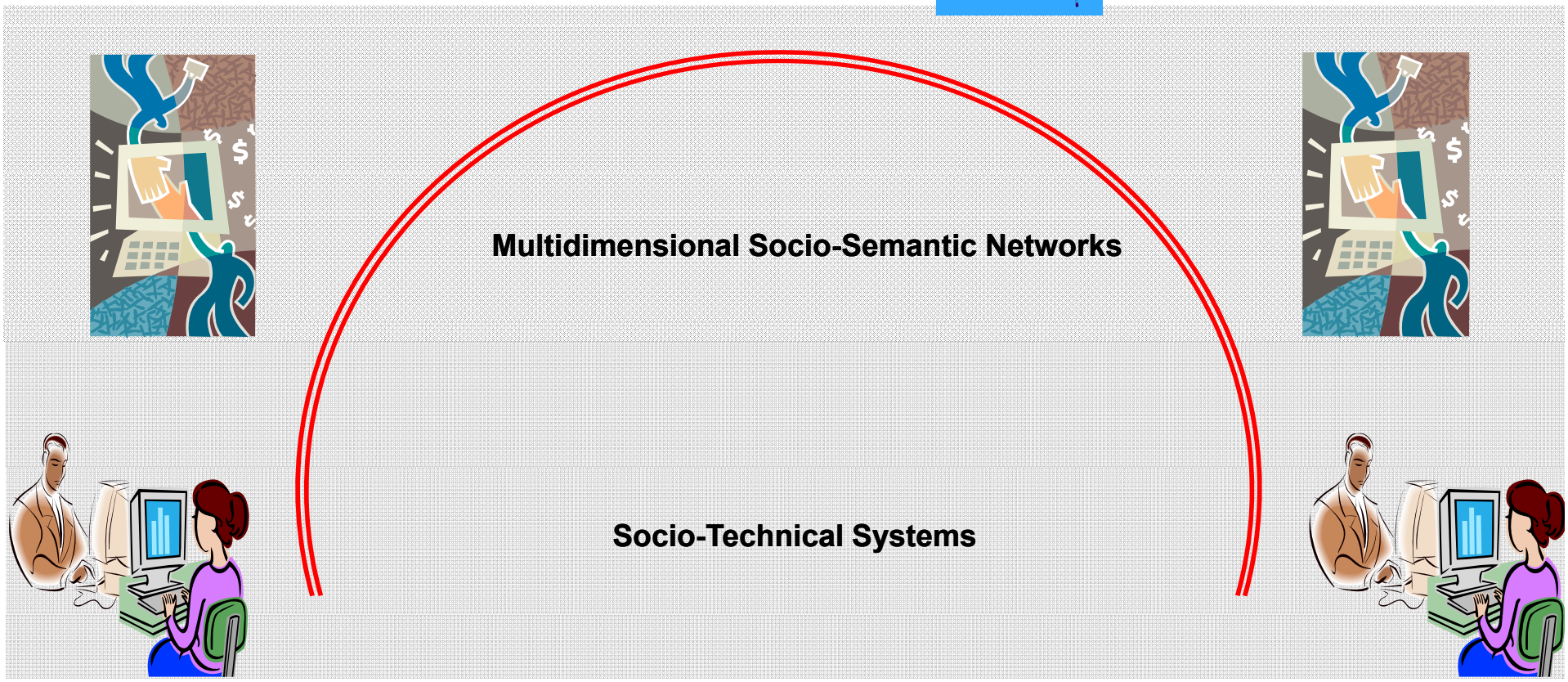
Networks and Graphs: Introduction

Examples of Networks

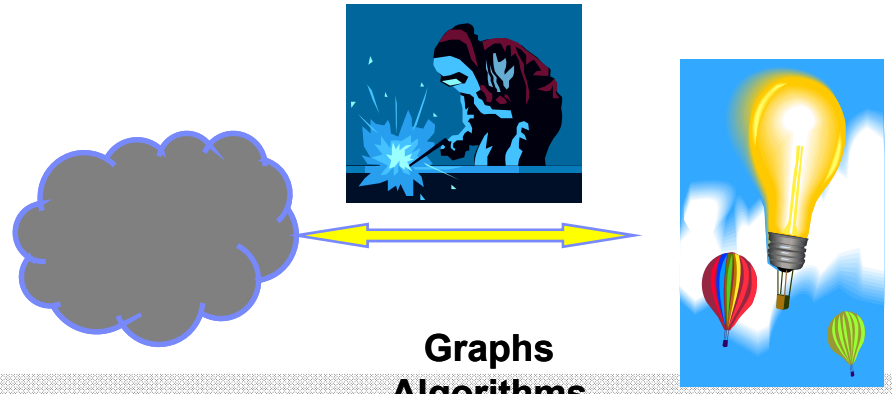
- Many natural and human-made networks:
  - Social networks (acquaintances, movie actors, email, trust, collaboration, citations, organizational management, …)
  - Communication (email, telephone, wireless)
  - Economy (trade networks, currency, industry, wealth, tourism,...)
  - Financial Market
  - Computer Science (software architecture, data sharing, spam filtering, circuits, ...)
  - Internet, WWW
  - Transportation (airports, roads and urban streets, ...)
  - Electric power transmission systems
  - Medicine, Neuroscience, Biomolecular Networks (Protein-protein Interaction Networks, Metabolic Networks, Genetic Networks)
  - Linguistics (Semantic Networks, Superficial Networks)
  - Earthquakes
  - Physics (Energy landscapes, Astrophysics, ...)
  - Chemistry
  - Mathematics
  - Climate networks
  - Security and Surveillance
  - Epidemic spreading

Three layers





Agenda

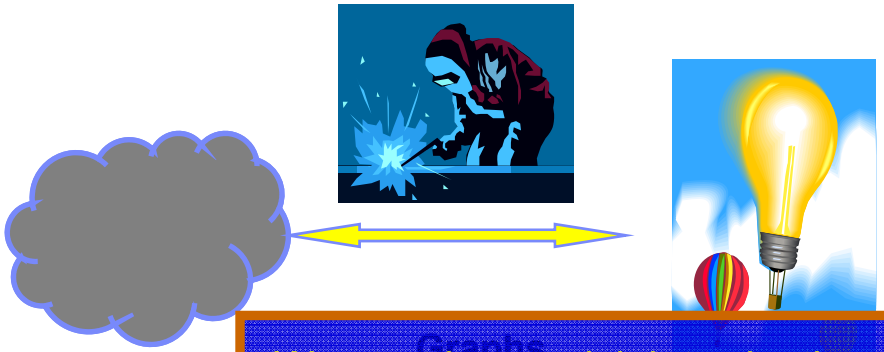


Graphs
Algorithms

Multidimensional Socio-Semantic Networks

Socio-Technical Systems
Social or Behavioral Networks

Agenda



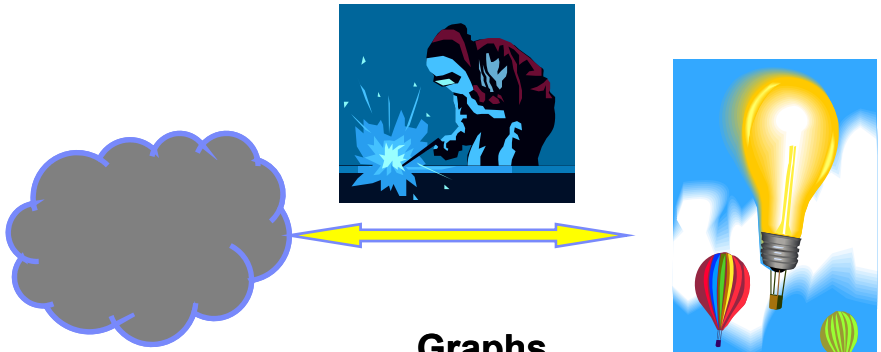
We start by a quick introduction to socio-technical systems like Del.icio.us, Flickr, Digg, etc

We identify key challenges in harvesting social knowledge from these systems:
- community identification,
- ontology generation,
- user and document recommendation, etc.

and outline solutions to help people to discover and share content, to manage workflow and social connections.

The focus of our attention will be on the based on graph-based methods for mining of networks of socio-semantic connections

Agenda (Cont.)



Graphs

Algorithms

Multidimensional Social Semantic Networks

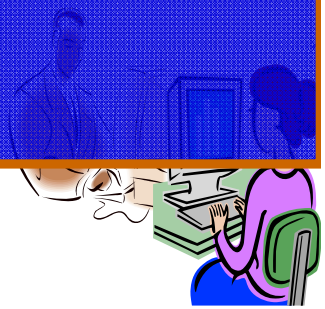
Our approach is similar to classical software approach - like Rational Unified Process:

We model real life phenomena by multidimensional networks. Network representation may hide important information. Moving towards higher levels of abstraction, we need to preserve

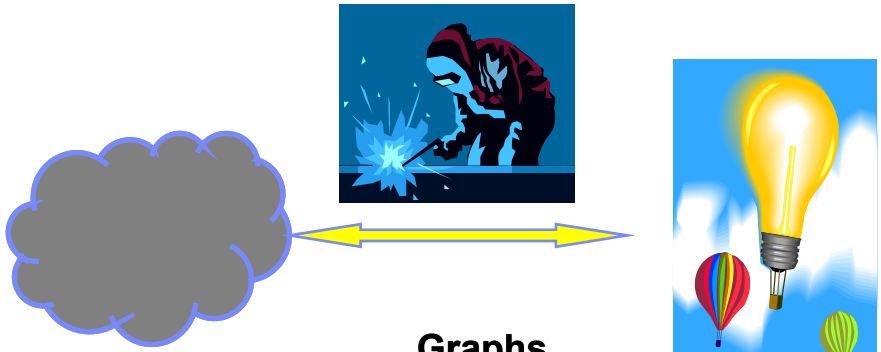
- -Semantics
- -Social,
- -Eevocative and
- -Workflow Management

aspects of real use of real systems

Socio-Technical Systems



Agenda (Cont.)



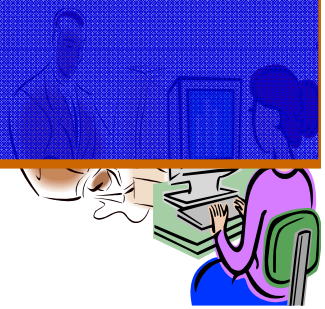
Graphs
Algorithms

Multidimensional Socio-Semantic Networks

We model networks by graphs
(similar concept but resides in a different domain).

In addition, in our approach:.
- Network - model of raw data and processes in real word systems
- Graph - is a task specific representation of a network

Socio-Technical Systems



Avenues to deep socio-semantic analytics and the possibility of high-quality functionalities for socio-technical systems (like recommending people to invite into your social network) hinge on the availability of engines which are able

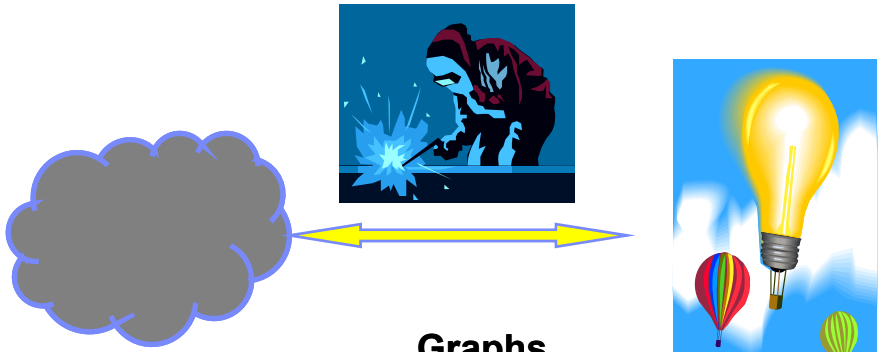
- to provide hidden knowledge discovery
  - like discovering a new relation in a network - that based on the strength of multiple connectivity between the nodes of a social network one can conclude that Dr. Jekyll is related to Mr. Hide)
- and provide ad hoc generalisation across dimensions.
  - For instance, the ability to detect that a particular person might serve as an representative of a community or as an expert on a particular topic (the example of such generalisation is the expression frequently attributed to Louis XIV *"I am the State"*

Graphs
Algorithms

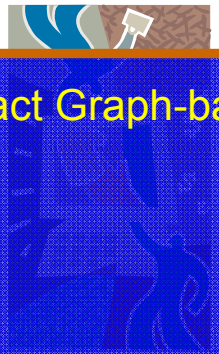
Socio-Semantic Networks

Socio-Technical Systems

Agenda (Cont.)



Graphs
Algorithms

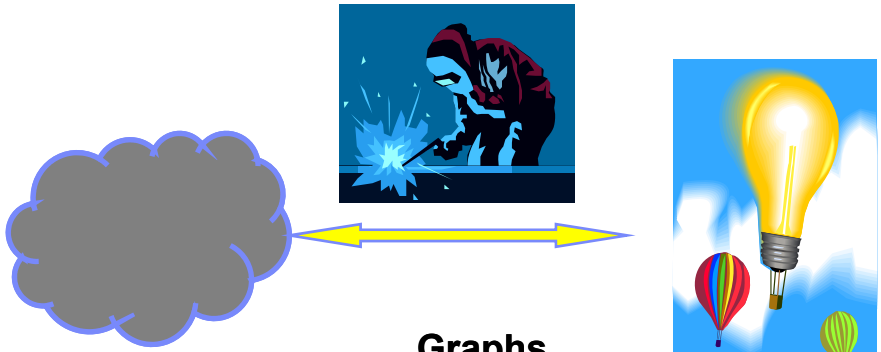


Abstract Graph-based Methods

Multidimensional Socio-Semantic Networks

Socio-Technical Systems

Agenda (Cont.)



Graphs
Algorithms

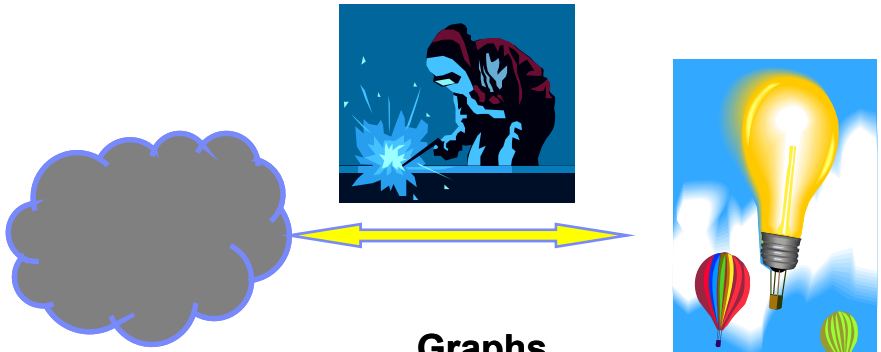
Graphs provide a powerful abstraction of the structure and dynamics of diverse kinds of interpersonal or people-to-technology interactions.

Multidimensional Socio-Semantic Networks

Through our tutorial we'll demonstrate how various graph-mining techniques (including soft clustering and fuzzy inferencing) can be used to reveal implicit social structures (like to detect ad hoc communities of common interests or practice).

Socio-Technical Systems

Agenda (Cont.)



Graphs
Algorithms

Mathematics meets reality:

:

We will briefly outline how to use explicit and implicit social structures to improve the consumability of social knowledge

antic Networks

Socio-Tech toms

Network and Graph Theory Glossary

Network and Graph Theory Glossary (Introduction)

- Graph theory: Vertices-(Arcs or Edges)
- Comp. Sci.: Nodes-Links ; States-Transitions
- Network Analysis as a branch of sociology and mathematics: Actors-(Ties, Links)

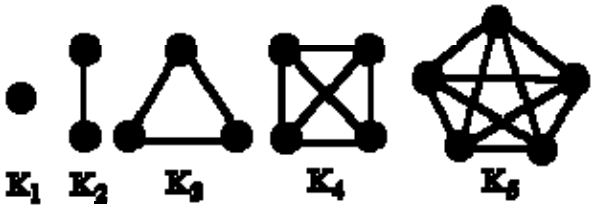
Network and Graph Theory Glossary

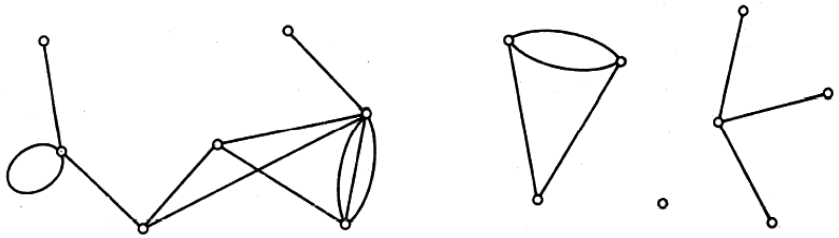
- Matrix: adjacency matrix - V×V matrix; two vertices are adjacent if they are connected by an edge
- Matrix: incidence matrix – E×V matrix
- Arc ~ “edge” in directed graphs
- Clique – SNA term ~ complete graph
- Complete graph - each vertex is connected to each of the others
- Connected - A graph is connected if there is a path connecting every pair of vertices. A graph that is not connected can be divided into connected components
- Degree - The degree (or valence) of a vertex is the number of edge ends at that vertex. In a digraph (directed graph) the degree is usually divided into the in-degree and the out-degree.
- Diameter – Graph Diameter is the length of the longest shortest paths
- Digraph - A digraph (or a directed graph) is a graph in which the edges are directed.
- Edge ~ arc, link
- Graph – *G:=(V,E)* *V*- vertices, *E* - edges
- Loop - A loop is an edge that connects a vertex to itself.
- Path - A path is a sequence of consecutive edges in a graph and the length of the path is the number of edges traversed.

Source: http://www.utm.edu/departments/math/graph/glossary.html

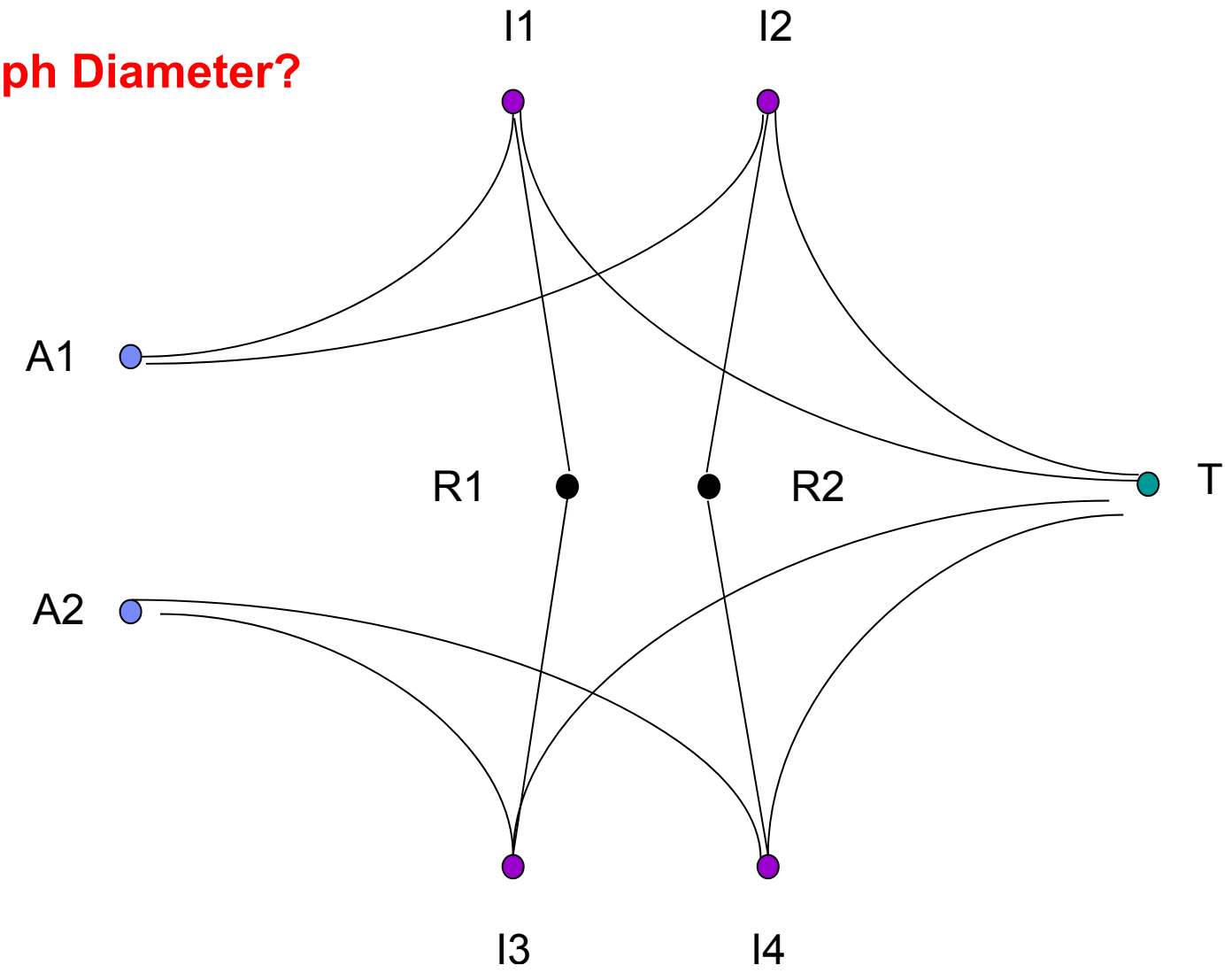
Glossary Illustrations

- Complete graph
- Connected/Disconnected Graphs

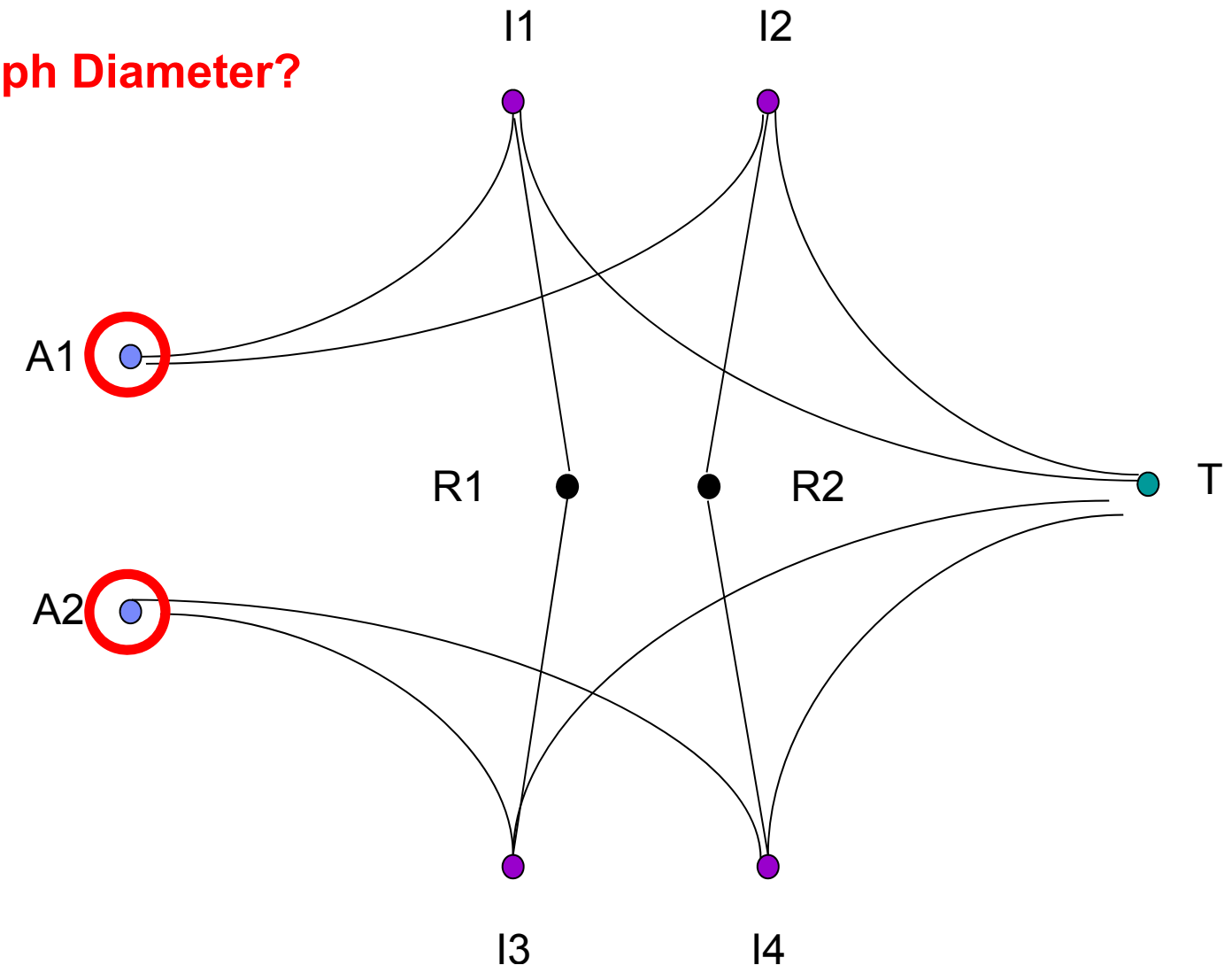




Graph Diameter?



Graph Diameter?



Examples of Graph Models: Two Mode Networks

- Two-mode networks consists of two sets of units; relation connects the two sets
  - Membership in institutions - people/institutions
  - Buying articles in the shop - consumers/articles
  - Co-authorship networks - authors/papers
- A corresponding graph is called bipartite graph
  - Edges connect only vertices from one to vertices from another set, inside sets there are no connections.

Example of Graph Models: Data → Matrix → Graph

- Market Basket Data

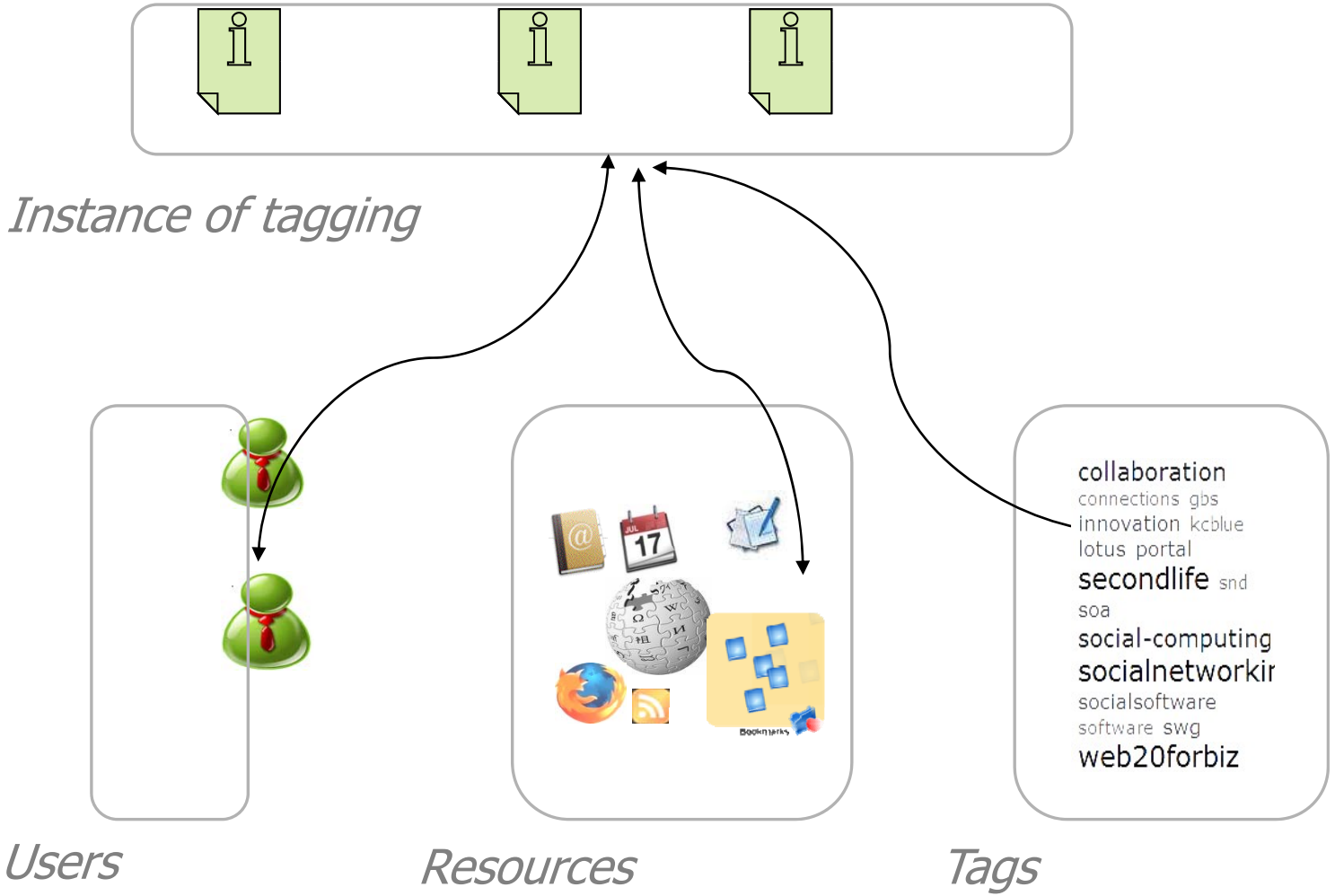
	Bread	Milk	Wisky	…
Jack	4	2	0	
Katie	3	10	0	
Conor	1	0	3	

Examples of Graph Models: Folksonomies: – Tripartite Hypergraph

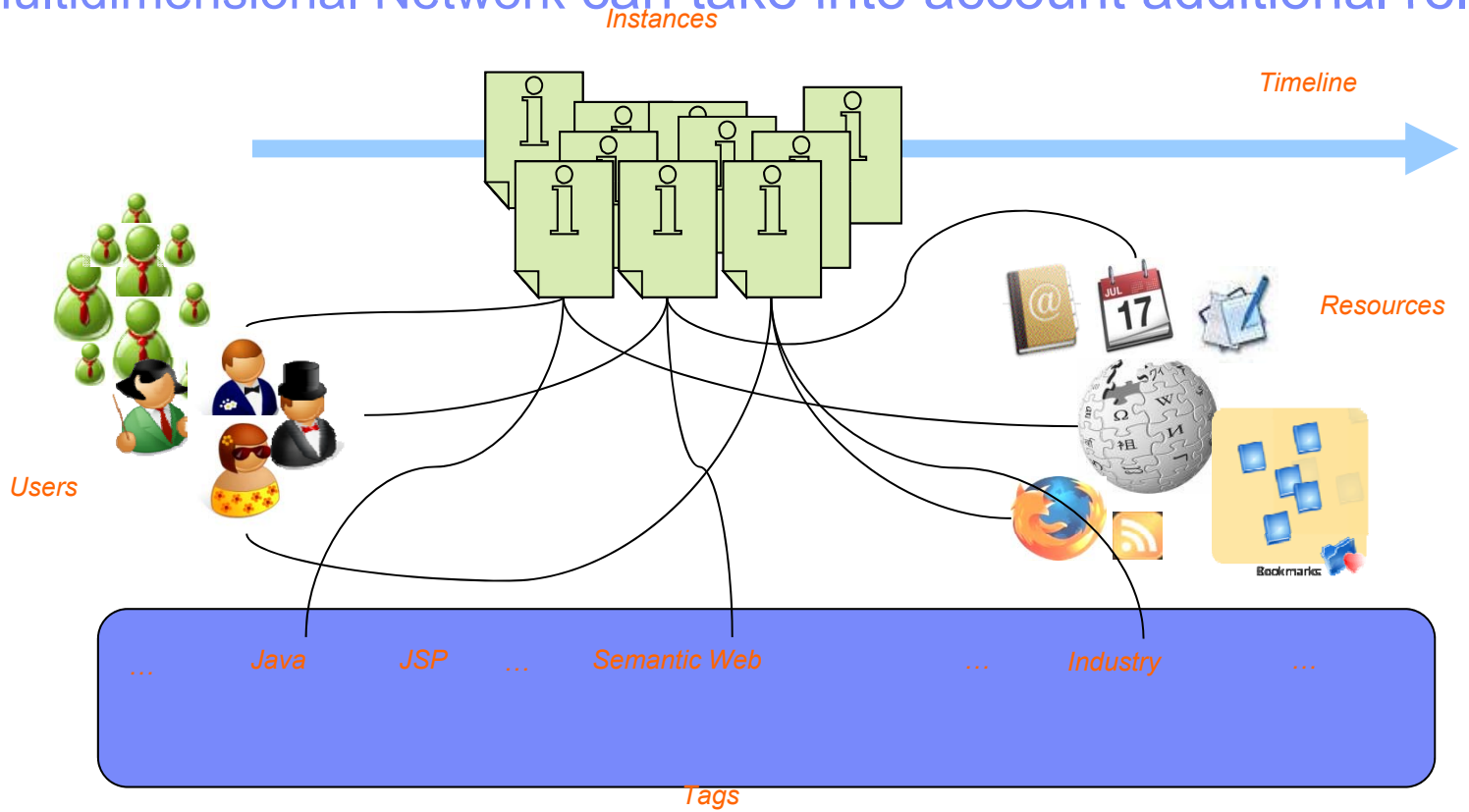
- In social bookmarking systems users describe bookmarks by keywords called tags. The structure behind these social systems, called folksonomies, can be viewed as a tripartite hypergraph of user, tag and resource nodes.

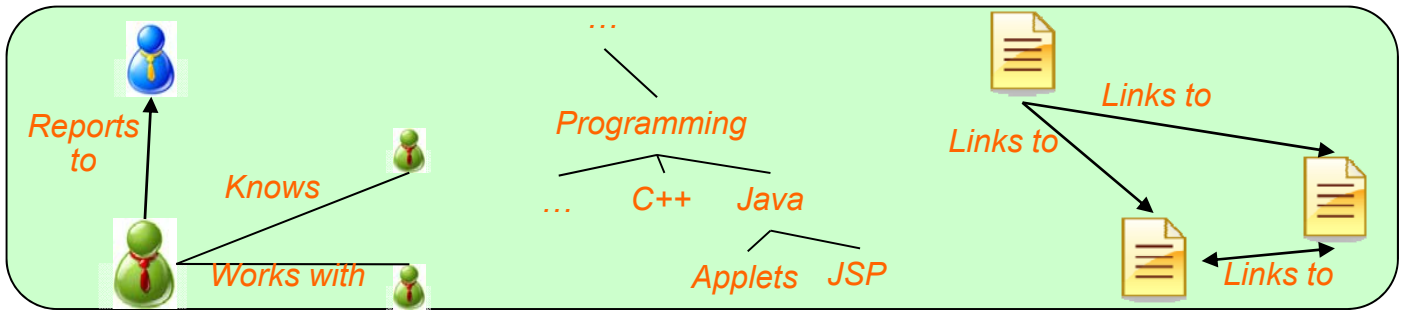
Source: Jäschke et al. "Logsonomy — A Search Engine Folksonomy" MediaICWSM 2008AAAI Press (2008)

Examples of Graph Models:
Folksonomies: – Multidimensional Network



Multidimensional Network can take into account additional relations





Terms brought by physicists

Percolation

- In physics, chemistry and materials science, percolation concerns the movement and filtering of fluids through porous materials.
  - coffee percolation, where the solvent is water, the permeable substance is the coffee grounds, and the soluble constituents are the chemical compounds that give coffee its color, taste, and aroma.
- During the last three decades, percolation theory, an extensive mathematical model of percolation, has brought new understanding and techniques to a broad range of topics
  - Social networks – information percolation

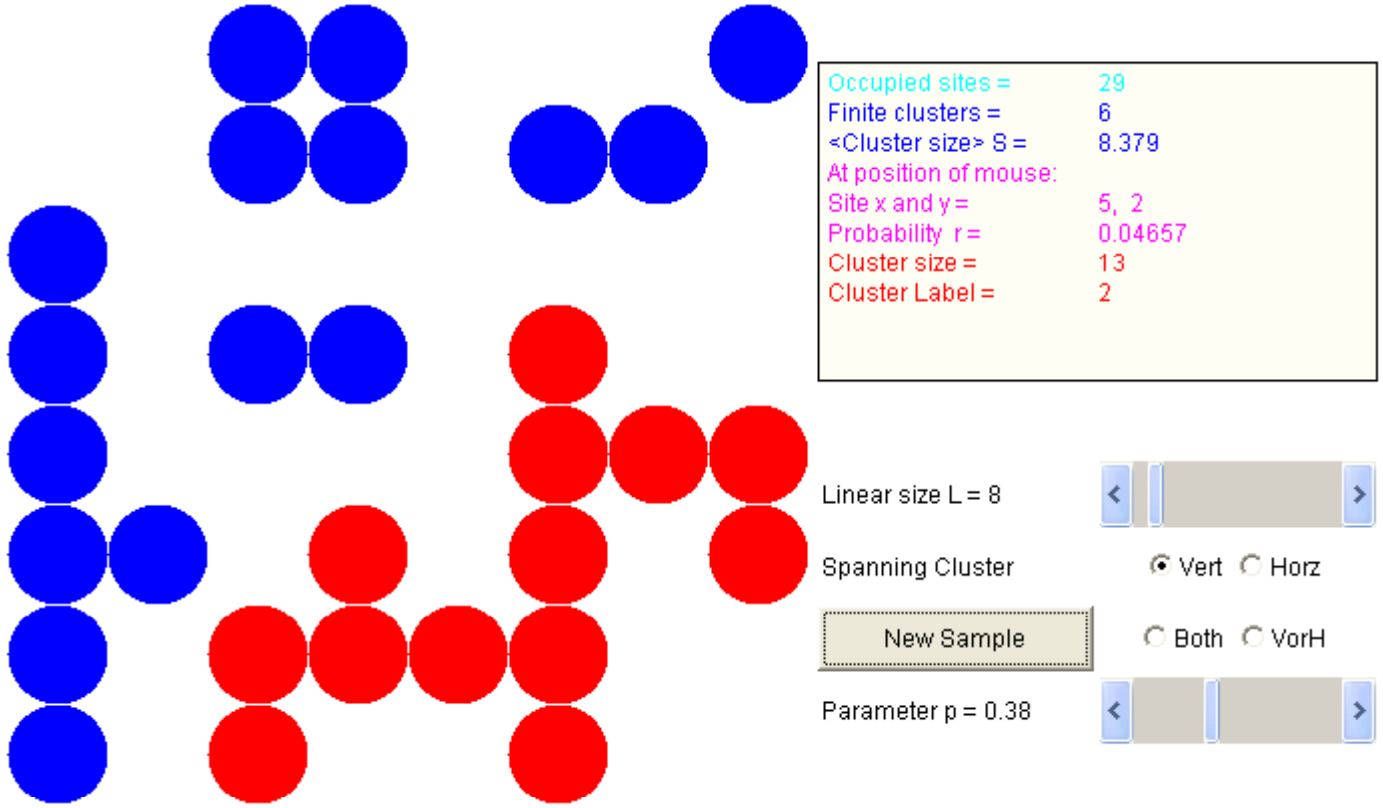
Percolation in physics:
Conductive connection between opposing sides of a square

- A square plate covered by metal shavings. Some of these metal shavings touch; others don't. A high concentration of metal shavings gives you a greated chance of a conductive connection between opposing sides of a square; lower concentration of metal shavings reduce that chance.
- What is the lowest possible amount of metal shavings that still provides a reliably connected network?

Shavings - thin pieces that have been cut from the surface of something, especially wood or cheese (macmillandictionary.com)

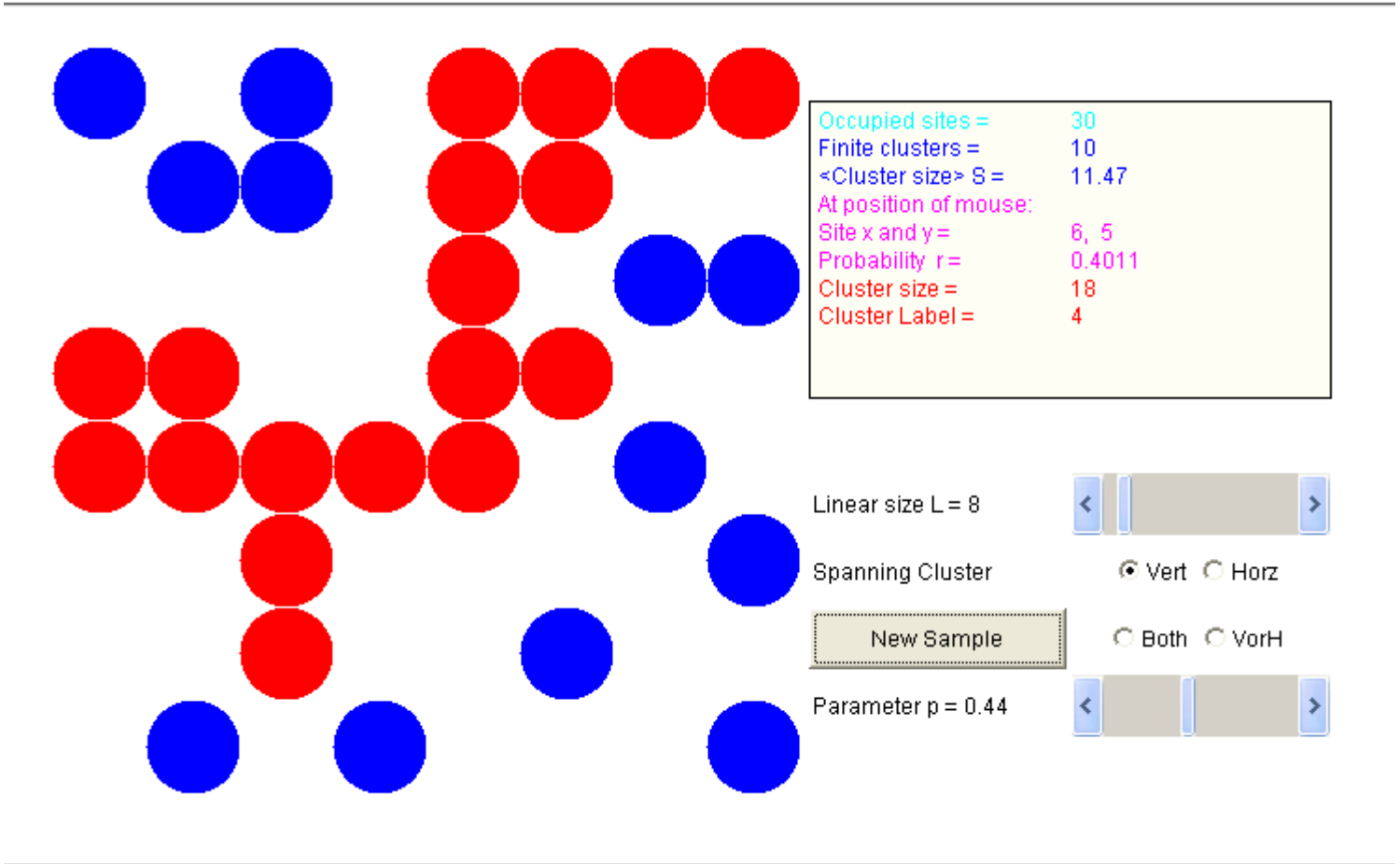
Percolation on two dimensional grid

Concentration ~ Number of occupied sites



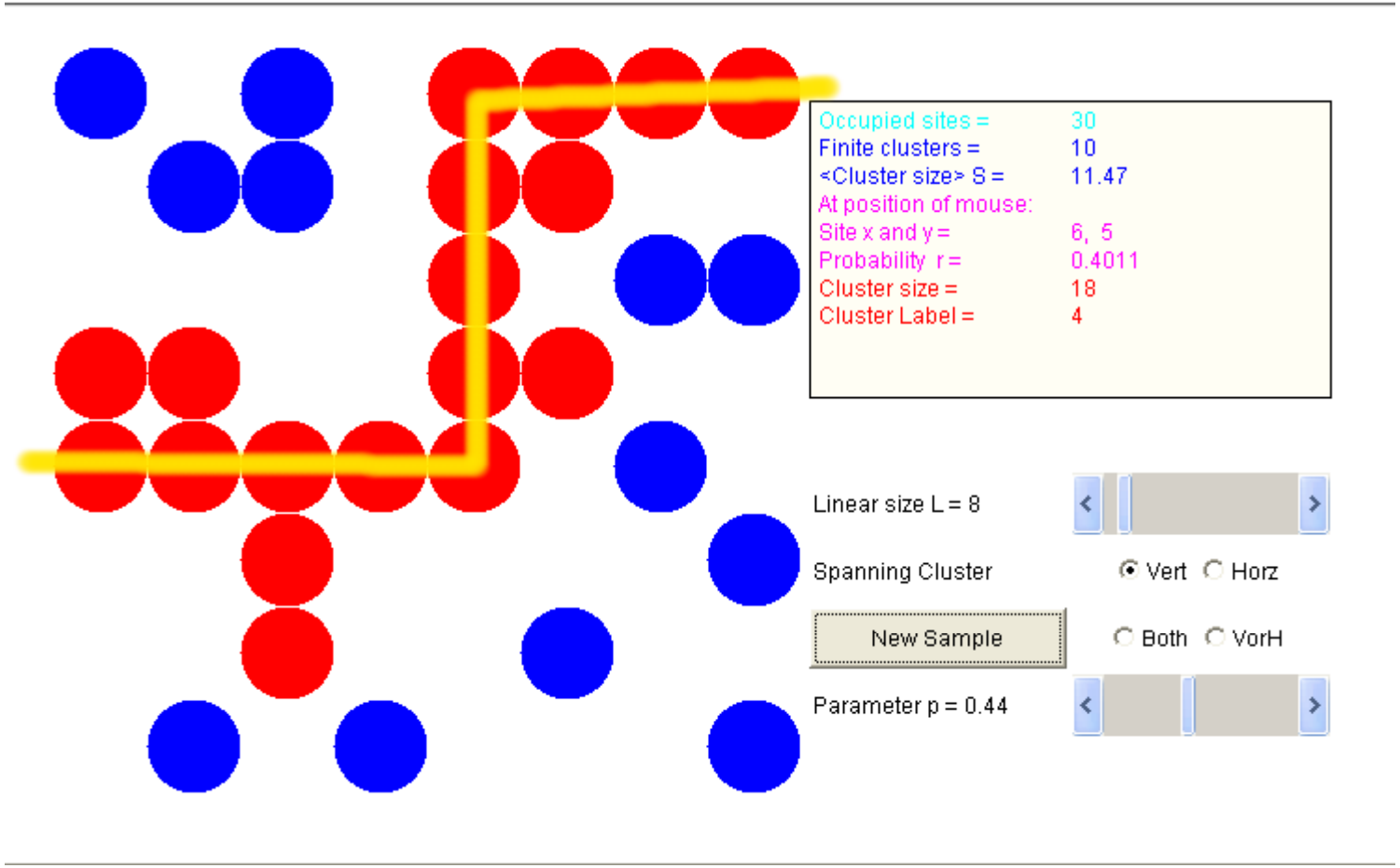
Source: http://www.physics.buffalo.edu/gonsalves/ComPhys_1998/Java/Percolation.html

Percolation on two dimensional grid



Source: http://www.physics.buffalo.edu/gonsalves/ComPhys_1998/Java/Percolation.html

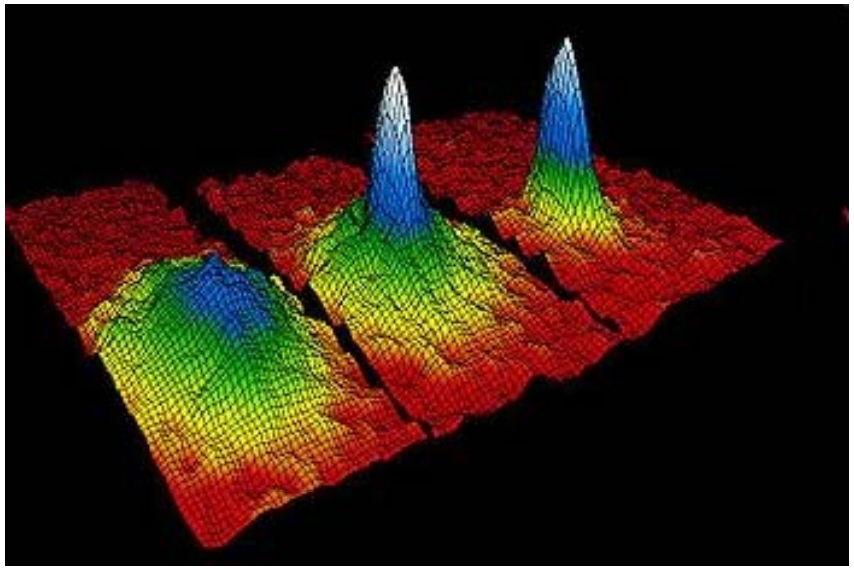
Percolation on two dimensional grid



Source: http://www.physics.buffalo.edu/gonsalves/ComPhys_1998/Java/Percolation.html

Bose–Einstein condensation: a network theory approach

- In physics, a Bose–Einstein condensate is a state of matter that occurs in certain gases at very low temperatures.
- A similar type of self-organization (condensation transition) can occur also in complex networks. In this context, a condensation phenomenon occurs when a distribution of a large number of elements in a large number of element classes becomes degenerate, i.e. instead of having an even distribution of elements in the classes, one class (or a few classes) become occupied by a finite fraction of all the elements of the system.
  - Condensation transitions occur in traffic jams, where long queues of cars are found, in wealth distribution models where a few people might have a finite fraction of all the wealth or in spin glass models. However, the condensation transition in these models cannot in general be mapped to a Bose-Einstein condensation.



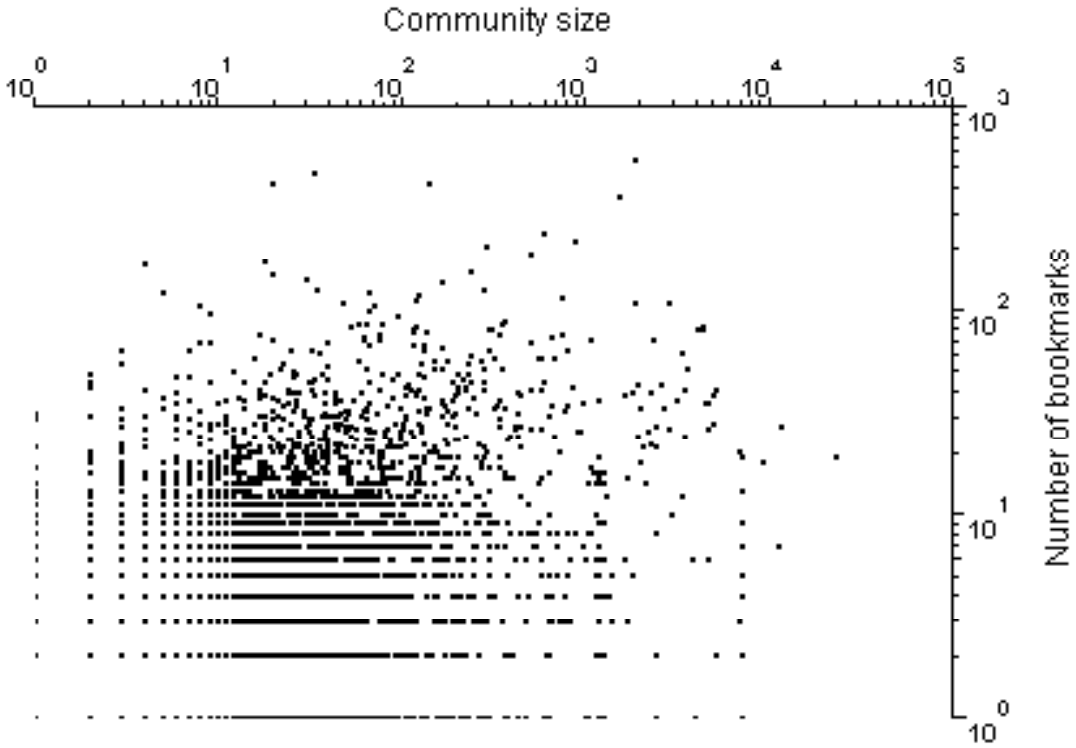
Bose–Einstein Condensation at 400, 200, and 50 nanokelvins. The peaks show that as the temperature goes down, more and more atoms "condense" to the same energy level.

Source: Wikipedia

Complex Systems, Emergency

- Complex System
  A complex system is a system composed of interconnected parts that as a whole exhibit one or more properties (behavior among the possible properties) not obvious from the properties of the individual parts.
- Emergency
  The complexity of many real-world systems frequently exceeds that predicted from knowledge of the individual elements and interactions between them.
  - The concept of emergence in the social sciences (as well as in physical and biological sciences) is an elusive one: "Hydrogen is a light, odourless gas, which, given enough time, turns into people." (sometimes attributed to Edward R.Harrison, Cosmologist)

Example of Emergency



- Bigger communities generate more activities? NO
  - There is no correlations between age/size and the level of user participation
  - For instance, this plot shows the absence of correlation between Community size and Number of bookmarks

The log-log plot of community size vs. number of bookmarks generated in the community. Each community is represented by a dot in the log-log scale: x=log(community size), y=log(number of bookmarks).
The cloud of the dots seems to be randomly and almost uniformly covering the rectangle. This shows that there is no dependency between community size and number of bookmarks produced in the community.

Figurese: Nevidomsky, A. and Troussov, A. "Structure and Dynamics of Enterprise 2.0 Communities", Web Science Conference 2010

Is it because “scale kills conversation”?

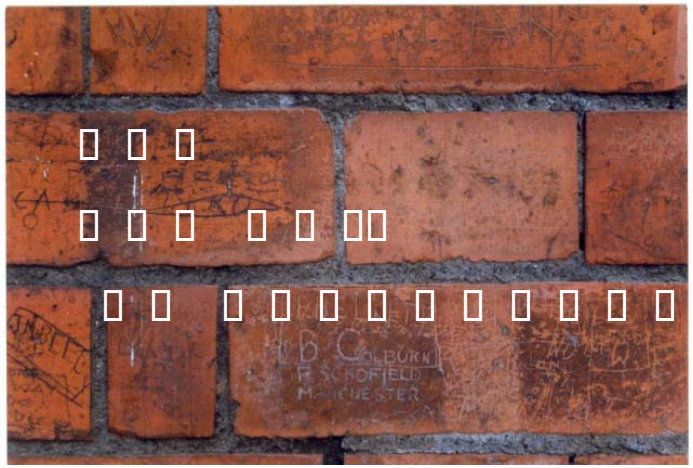
"And, finally, you have to find a way to spare the group from scale. Scale alone kills conversations, because conversations require dense two-way conversations”. - Clay Shirky (2003)

- NO or NOT QUITE
  - We quantify “the scale” of the group as the log of the number of members, and, correspondingly, the “conversations” as the scale of the user participation – i.e. log of the total number of artifacts related to user activity in it (like forum topics or bookmarks)
  - In our data
    “SCALE” HAS NOTHING TO DO WITH “CONVERSATIONS”
    while if the rule “scale kills the conversation” would be applicable, we would find positive or negative correlations between “scale” and “conversations”
  - The whole is greater than the sum of the parts. We can tentatively suggest, that the scale of user participation could be better explained by social and structural properties of communities, rather than by functioning of individuals. Homogeneity and heterogeneity of communities and graph-based structural properties of the social networks behind the online community (connectedness, distance, density, etc.) might have strong correlation with user participation.

Source: Clay Shirky A Group Is Its Own Worst Enemy. Keynote speech on Social Software at the O'Reilly Emerging Technology conference, Santa Clara, April 24, 2003.

Reductionism in XX century

Reductionism →



← **Self-organization**

Emergency

Alexander Troussov, Ph.D., IBM Dublin Software Lab

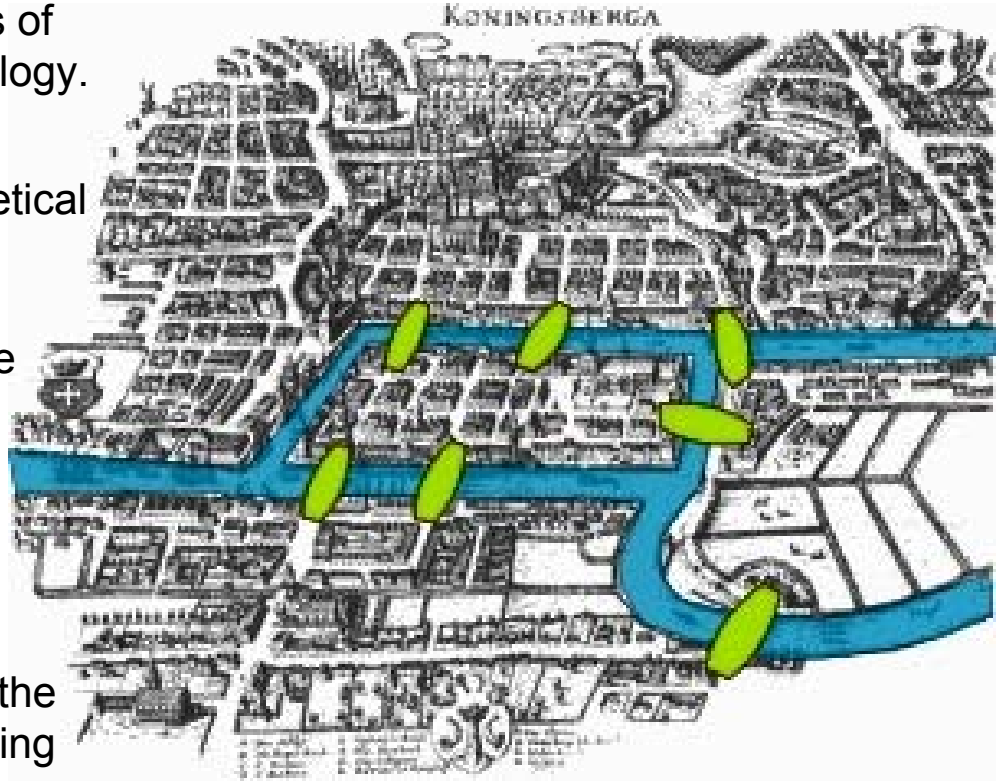
4th Russian Summer School in Information Retrieval, September 13-18, 2010, Voronezh

RuSSIR
Russian Summer School
in Information Retrieval

Networks and Graphs: Random vs. Real World Networks

Leonhard Euler: Seven Bridges of Königsberg

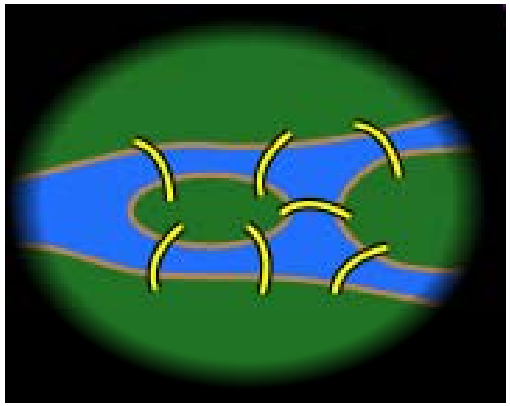
- Leonhard Euler in 1735 laid the foundations of graph theory and presaged the idea of topology.
- Theory of graphs has been useful for theoretical physics, economy, sociology and biology. However, most of such studies focused on **static graphs**, i.e. graphs whose structure remained fixed.
- Map of Königsberg in Euler's time showing the actual layout of the seven bridges, highlighting the river Pregel and the bridges



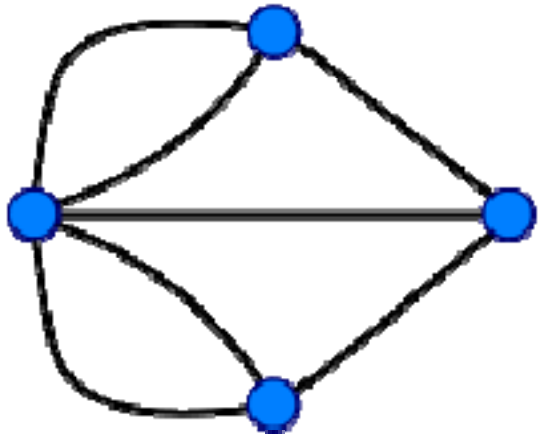
Source: Wikipedia

Leonhard Euler: Seven Bridges of Königsberg (Cont.)

- Leonhard Euler represented four land areas by nodes A, B, C, and D.
- The bridges – by links



- **Existence of the path** does not depend on our ingenuity to find it.
  It depends on …
- It depends on **a property of the GRAPH**



Source: "Linked", Wikipedia

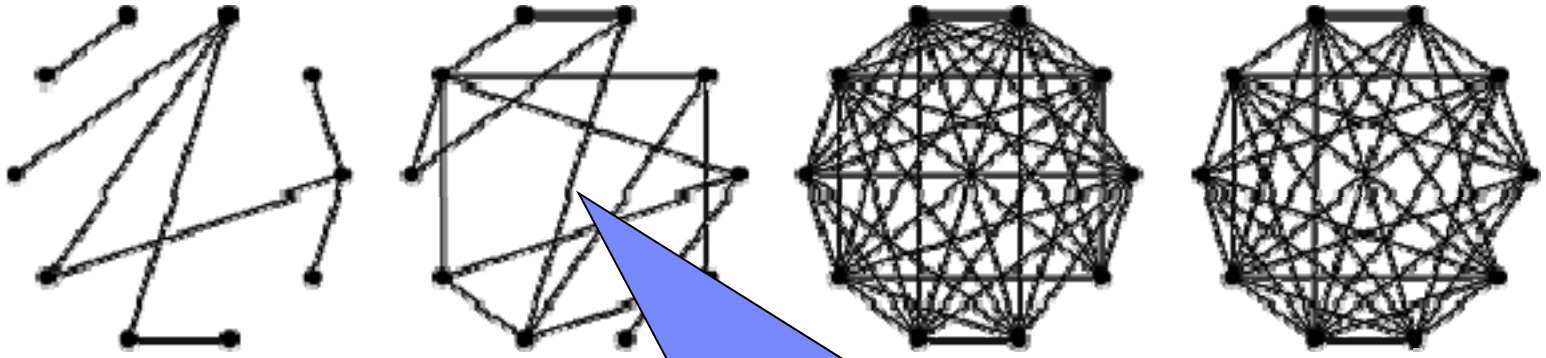
Erdős–Rényi – the Concept of Random Graphs

- Paul Erdös and Alfred Rényi were the first to introduce the concept of random graphs in 1959.
- In graph theory, **the Erdős–Rényi model**, named for Paul Erdős and Alfréd Rényi, is either of two models **for generating random graphs**, including one that sets an edge between each pair of nodes with equal probability, independently of the other edges.
  - a graph in which either a fixed number of edges are randomly distributed among all the pairs of a set of nodes, or, alternatively, a graph in which every pair have the same independent probability of being connected (Bollabás 2001).

- A.T. Erdős Number=5
  - A. V. Troussov coauthored with S.N. Gurbatov who coauthored with Uriel Frisch who coauthored with Ciprian Foias who coauthored with Jacques Dixmier who coauthored with Paul Erdős

Random graphs

A random graph is a graph in which properties such as the number of nodes, edges, and connections between them are determined in some random way.



Edge probab

Erdös and Rényi simp
fixed number of verti
by edges.

Start with $N = 10$ isolated nodes, then connect every pair of nodes with probability $p = 0.3$

Erdős–Rényi model

- *G(n, p)* - undirected Erdos-Renyi graph.
- Every edge is formed with probability p ∈ (0, 1) (independently of the presense of other edges).
- Let *Iij* ∈ {0,1} be a Bernoulli random variable indicating the presence of edge {*i,j*}
- Random variables *Iij* are independent and I_{ij} = 1 with probability *p*, I_{ij} = 0 with probability 1-*p*
- E[number of edges] = E[∑ *Iij*] = (*n*(*n*-1)/2)*p*
- Using weak law of large numbers, we have for all *α* > 0

$$\mathbb{P}\left(\left|\sum I_{ij} - \frac{n(n-1)}{2}p\right| \geq \alpha\frac{n(n-1)}{2}\right) \to 0,$$

- as *n* → ∞. Hence, with this random graph model, the number of edges is a random variable tightly concentrated around its mean for large *n*

Phase Transition examples

Exhibiting phase transitions was one of the main contributions of Erdosh and Renyi 1959

Giant cluster everges on average number 1 for links (large fraction of all nodes is connected)

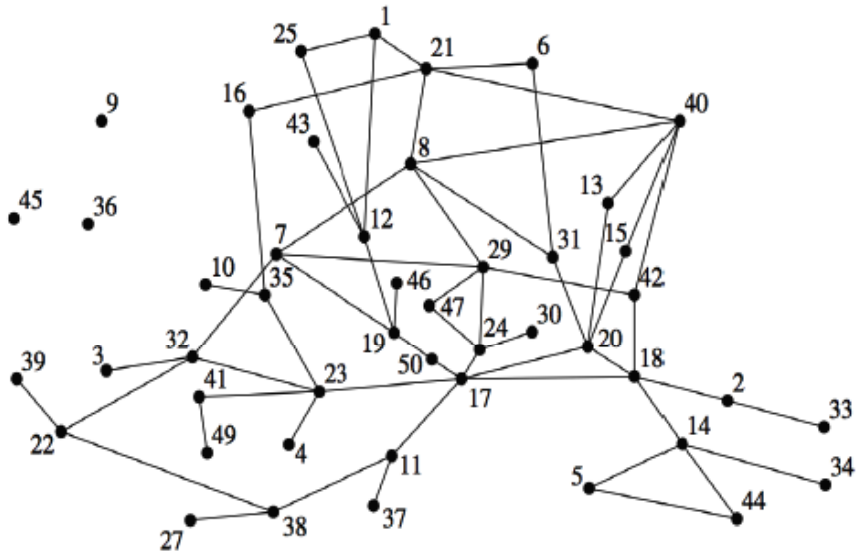


Figure: Emergence of a giant component: a random network on 50 nodes with $p = 0.05$.

Source: Daron Acemoglu "Erdos-Renyi Graphs and Phase Transitions"

Phase transition examples

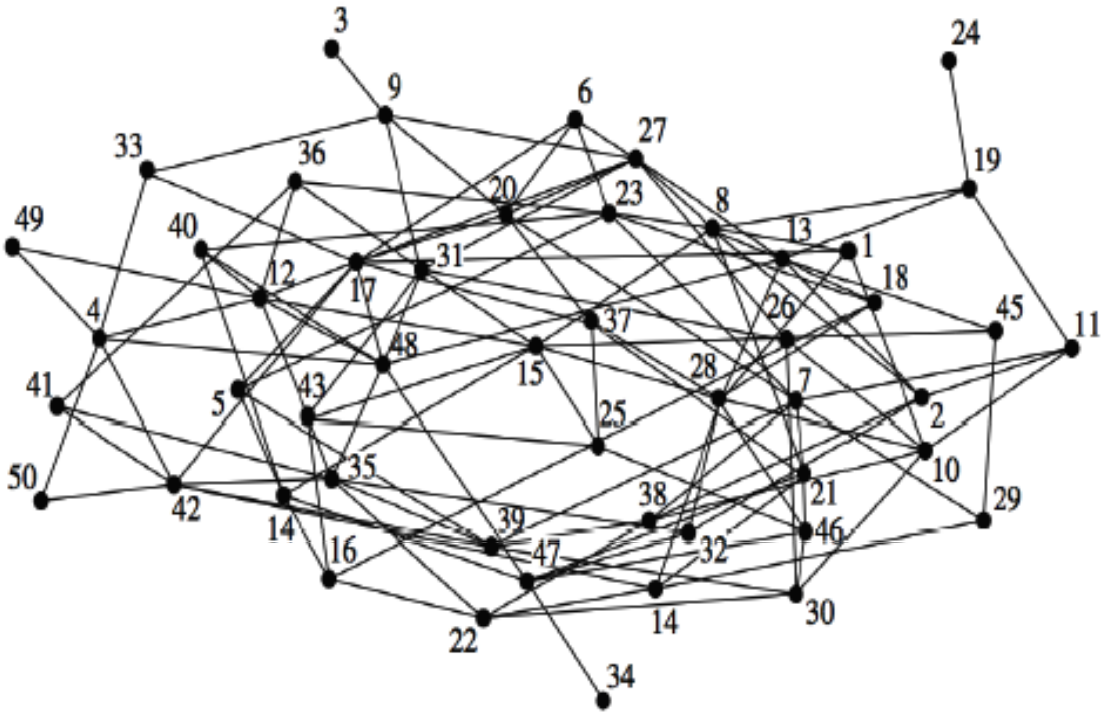


Figure: Emergence of connectedness: a random network on 50 nodes with $p = 0.10$.

Source: Daron Acemoglu "Erdos-Renyi Graphs and Phase Transitions"

Phase Transition, Threshold Function

- Erdos-Renyi model is completely specified by the link formation probability *p*(n).
- For a given property *A* (e.g. connectivity), we define a threshold function *t(n)* as a function that satisfies:

$$\mathbf{P}(\text{property } A) \to 0 \quad \text{if } p(n)/t(n) \to 0$$

and

$$\mathbf{P}(\text{property } A) \to 1 \quad \text{if } p(n)/t(n) \to \infty$$

This definition makes sense for “monotone or increasing properties," i.e., properties such that if a given network satisfies it, any supernetwork (in the sense of set inclusion) satisfies it.

- When such a threshold function exists, we say that a phase transition occurs at that threshold.

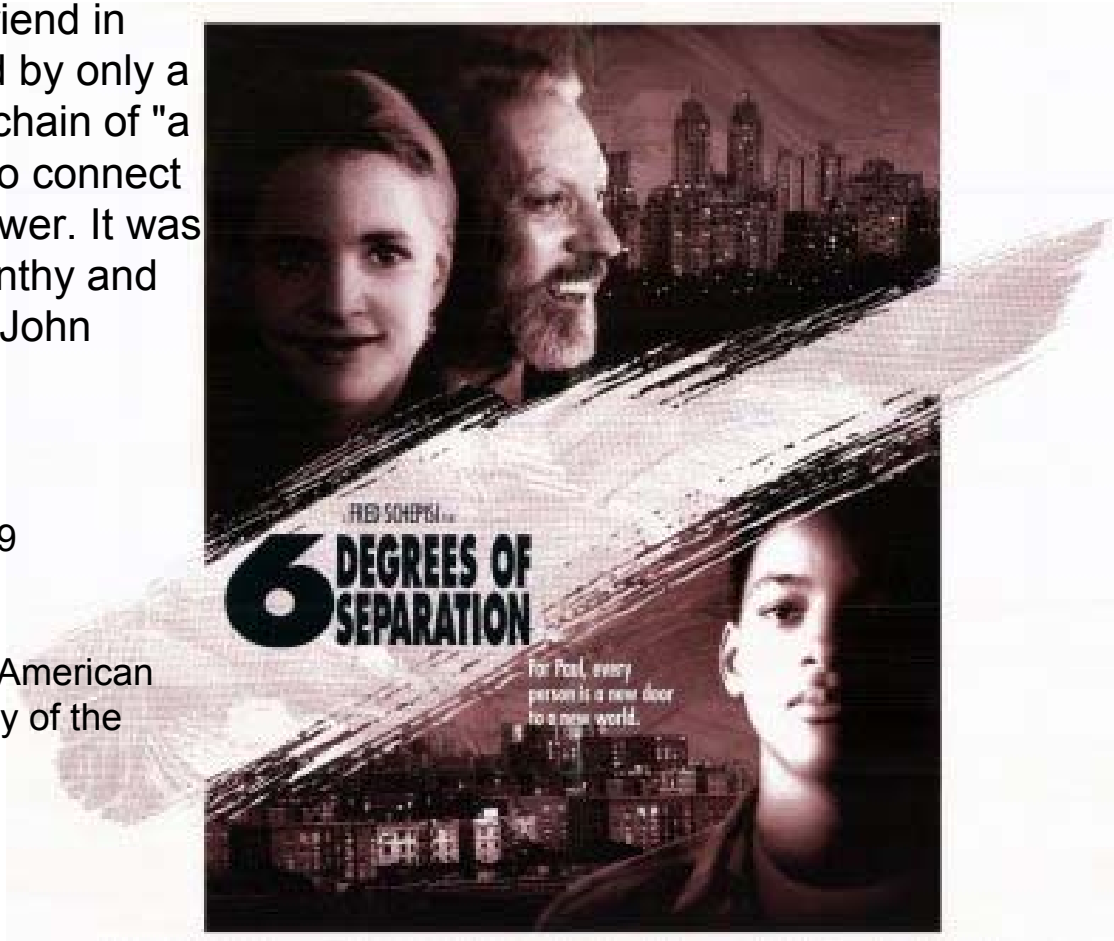
Erdős–Rényi model – Mathematics meets Reality

- Erdős–Rényi model may be inappropriate for modeling real-life phenomena.
  - "This **mathematically tractable but completely unrealistic model** has been used as a baseline in epidemiology and other fields. Random networks completely lack structure: there is no tendency to form clusters (cliques); actors to differ in their propensities of contacts; there are no tendencies for centralization or transitivity. In fact, no conceivable bias exists in random graphs – **no leadership, no homophily of choice – no nothing**. Compared to actual social networks, random graphs have a low level of clustering (no cliquing), small differences in degree among the vertices, and short distances between vertices. In exchange for this unrealistic simplicity we can calculate other important features of networks such as the distribution of component sizes and average distances between nodes."
    The Invasion of the Physicists: Review of Duncan J. Watts, Six Degrees, and Albert-László Barábasi, Linked

    Homophily (i.e., love of the same) is the tendency of individuals to associate and bond with similar others. The presence of homophily has been discovered in a vast array of network studies.
- Despite the formalism and comprehensiveness of the theoretical results obtained by Erdos and collaborators, random networks ultimately proved not to be good models for natural structures and phenomena.
  - **Heterogeneous structuring, not the relative uniformity and simplicity of networks, is the rule in Nature (?)**
- For popular modeling alternatives, see Barabási–Albert model and Watts and Strogatz model.

Six Degrees of Separation

- When two people don't have a friend in common, they are still separated by only a short chain of intermediaries. A chain of "a friend of a friend" can be made to connect any two people in six steps or fewer. It was originally set out by Frigyes Karinthy and popularized by a play written by John Guare.



- Karynthy: “Láncszemek” (Chain) 1929

- Six Degrees of Separation is a 1993 American film adaptation of the John Guare play of the same title,

Source: Wikipedia

Stanley Milgram's small world experiment

- The small world experiment comprised several experiments conducted by Stanley Milgram (1967)
  - 296 people were asked to get a letter to a target person in Boston
  - letter could only be passed along a chain of firs-name acquaintances
  - 29% letters arrived
  - they took on average of 6.2 steps to get there (although some of this colleagues thought that it will take 100)
- In Graph terms - path lengths
  - Diameter (longest shortest path): $max\ d_{ij}$
  - Mean geodesic (shortest) distance

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij} \qquad \ell^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1}$$

The Invasion of the Physicists

- Duncan Watts and Albert-László Barábasi are both physicists who have recently crashed the world of social networks, arousing some resentment in the process. Both have made a splash in the wider scientific community, as attested by their publications in high status science journals (Science, Nature). Both have analyzed some of the same very large networks (for example, the internet). Both use models from physics - Bose-Einstein condensation, percolation, and so on. Both have recently written scientific best-sellers: Six Degrees ranks 2,547 on the Amazon list, while Linked ranks 4,003. These similarities, however, obscure profound and important differences between the two models they initiated. Watts and Barábasi had different purposes in creating their models, and the models are applicable in different situations.
  The Invasion of the Physicists: Review of Duncan J. Watts, Six Degrees, and Albert-László Barábasi, Linked

Random Networks and Milgram's experiments

- Watts:
  - Defects of random network models: Random networks are not clustered enough.
  - Explanation of Milgram's experiments: the random connectivity is imposed on a locally clustered social world.
- Barábasi
  - Defects of random network models: Random networks do not have hubs.
  - Explanation of Milgram's experiments:
    the distance-shortening effect of hubs (intuitively, the distances between the nodes slowly grow as the network grows like log n, but with the formation of hubs the distances between nodes actually slowly decrease

Four properties found in many real world networks

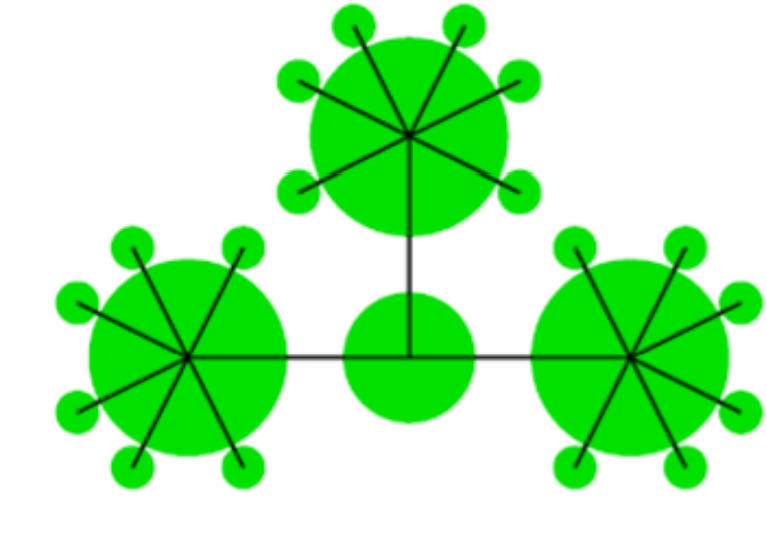
- Properties which seem to be common to many networks
  - the small-world property
  - power-law degree distributions,
  - and network transitivity.
  - and community structures

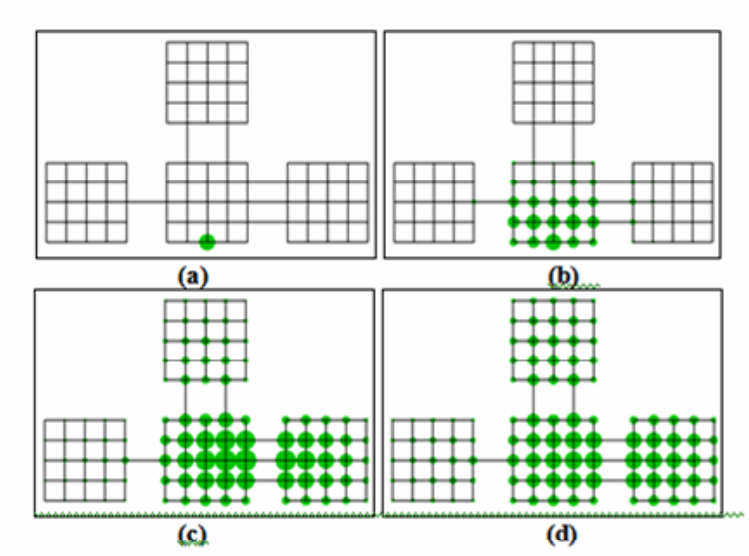
Four properties found in many real world networks

- Properties which seem to be common to many networks
  - Small-world property
    - Frequently, this term stands for short distances
    - More accurate definition:
      When densities are high and distances are short the network is declared to be a “small world.”
      Networks with a high degree of clustering may or may not be small worlds, depending on the patterns of weak distant ties in the network
  - Power-law degree distributions
    - Hubs and heavy tailed distribution
      There is small number of “Hubs”
      Coexisting with a large number of “less-significant” nodes
      Coexisting with larger number of “much-less” significant” nodes etc.
  - Network Transitivity
    - A friend of a friend is frequently also a friend
  - Property of Community Structure
    - in which network nodes are joined together in tightly-knit groups between which there are only looser connections
      Communities and dense sub-graphs are in abundance, the network is locally dense, globally sparse

“Small World” – Duncan Watts

- Duncan Watts: Based on Stanley Milgram’s “small world” experiment, real social networks have two (incompatible?) features – a high degree of clustering and relatively short path lengths.
- it is easy to design such networks
  - For instance, a set of disjoint cliques each one is connected by one tie to one central position would have both high density and short distances between nodes





Figurese: Troussov et al. "Vectorised Spreading Activation Algorithm for Centrality Measurement". MITACS FP-Nets Workshop on Social Networks, Vancouver, Canada. August 9-13 2010

“Small World” – Duncan Watts (Cont.)

- But since world-wide networks are not designed but emerge, Watts’s next important contribution is to show that there is a very simple uncoordinated process requiring no overall designer that produces a small world.
  - Watts starts with a model with only one of the properties of a small world, a network in which average path length is high but local density is also high. In one of his simulations (the simplest) this beginning model is a lattice in which individuals are connected to all neighbors within a certain distance. The microscopic process that creates a small world is random rewiring of this network. Each edge for each node is rewired with an independent probability β. If the edge of a node is rewired then another node is randomly selected as the other end of the edge. Watts shows that for small values of β the density of the network remains high but average distances between nodes approaches that of random graphs.
  - Watts’s goal is primarily conceptual. He defines a small world and shows that it can be generated without central planning. The model (a lattice) and the process (random rewiring) are completely and utterly unrealistic idealizations. No social network is a lattice and no network is randomly rewired in the way that Watts specifies. Rather, the lattice is an idealized instance of a network with local clustering, and random rewiring is an idealization of some process that perturbs this network and produces connections between nodes that were previously quite distant. A “test” of Watts’s model involves measuring average distances and local densities. When densities are high and distances are short the network is declared to be a “small world.” There is no claim that the examined network is a randomly rewired lattice.

Source: The Invasion of the Physicists: Review of Duncan J. Watts, Six Degrees, and Albert-László Barábasi, Linked

Scale Free - Barábasi

- Barábasi proceeds in quite a different direction, although he too uses random networks as a beginning touchstone. For Watts, random networks lack clustering, but for Barábasi the defect of random networks is that they have too flat a distribution of nodal degree. In a random network the distribution of degree follows a Poisson (or binomial) distribution. Consequently, most nodes have nearly the average degree and there is little variability. Many real networks, on the other hand, appear to have hubs – positions with many more connections than would occur in a random network. Instead of having the Poisson distribution, degree in these networks appears to follow a power law. These distributions have "fat tails," with a greater variance than for a Poisson distribution (or even a theoretically infinite variance). Consequently, they are more centralized than random networks and mean path lengths are even shorter than for random networks of the same average degree. Barábasi calls these networks scale-free because the average degree does not characterize most vertices.

Scale Free - Barábasi (Cont.)

- Barábasi's contribution is to suggest a plausible dynamic mechanism for why networks should have this scale-free shape: preferential attachment. A distribution with degree distributed approximately according to a power law will emerge from an initial small network if new nodes are continually being added that form attachments to existing nodes with probabilities proportional to the existing nodes' degree. In other words, the entering nodes tend to connect to the nodes that already have the most connections. This is a plausible scenario for many large networks. New web sites are most likely to link to popular and well-used sites. New members of a group may gravitate to its most popular members.

Power Low and Scale Free Networks - Barábasi

- Power law describes relations between the objects in the network, it describes scale invariance found in many natural phenomena (including physics, biology, sociology, economy and linguistics)

Scale Free Networks

- In a random network the distribution of degree follows a Poisson (or binomial) distribution. Consequently, most nodes have nearly the average degree and there is little variability.
- Many real networks, on the other hand, appear to have hubs – positions with many more connections than would occur in a random network. Instead of having the Poisson distribution, degree in these networks appears to follow a power law. These distributions have “fat tails,” with a greater variance than for a Poisson distribution. Consequently, they are more centralized than random networks and mean path lengths are even shorter than for random networks of the same average degree.
- Such networks are called scale-free because the average degree does not characterize most vertices.
- Plausible dynamic mechanism for why networks should have this scale-free shape: preferential attachments model
  - rich get richer - a new node (like new web site) links to an existing node (popular sites like Wikipedia) with probability proportional to its degree (how many links Wikipedia already has)

Power Low and Scale Free Networks

- Power law describes relations between the objects in the network, it describes scale invariance found in many natural phenomena (including physics, biology, sociology, economy and linguistics)

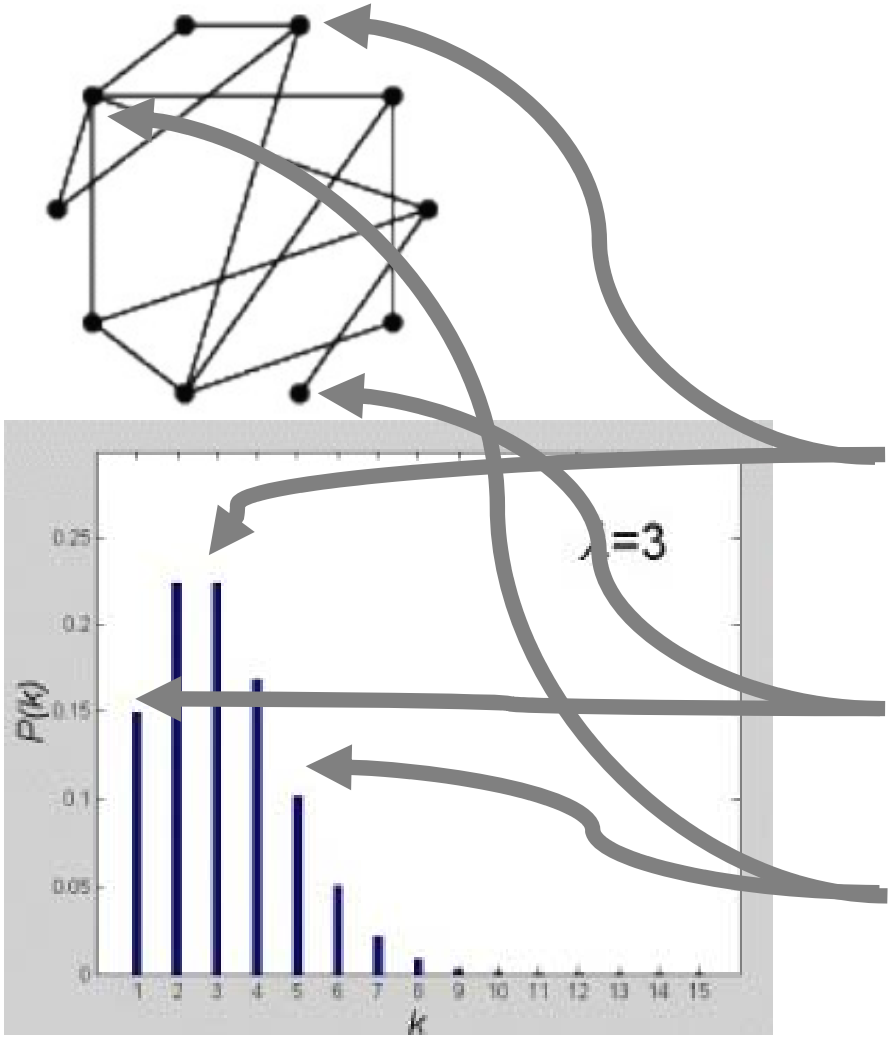
Scale Free Networks

- In a random network the distribution of degree follows a Poisson (or binomial) distribution. Consequently, most nodes have nearly the average degree and there is little variability. “Random universe is dominated by averages”
- Many real networks, on the other hand, appear to have hubs – positions with many more connections than would occur in a random network. Instead of having the Poisson distribution, degree in these networks appears to follow a power law. These distributions have “fat tails,” with a greater variance than for a Poisson distribution. Consequently, they are more centralized than random networks and mean path lengths are even shorter than for random networks of the same average degree.
- Such networks are called scale-free because the average degree does not characterize most vertices.
- Plausible dynamic mechanism for why networks should have this scale-free shape: preferential attachments model
  - rich get richer - a new node (like new web site) links to an existing node (popular sites like Wikipedia) with probability proportional to its degree (how many links Wikipedia already has)

Power Law: Introduction

- Most people are five- and six-feet-tall
  - seven- or eight-feet-tall, and three- or four-feet-tall individuals are rather rare
- $\Rightarrow$ There is an average height of adult humans
- However, not the all the distributions have a tendency to cluster around particular values:
  - Gutenberg-Richter: there are no "average" earthquakes!
  - Average number of visitors per a web-site?? Useless.
  - Pareto law in economics: there are a few multi-billionaires, but most people make only a modest income; "average" income is deceptive.

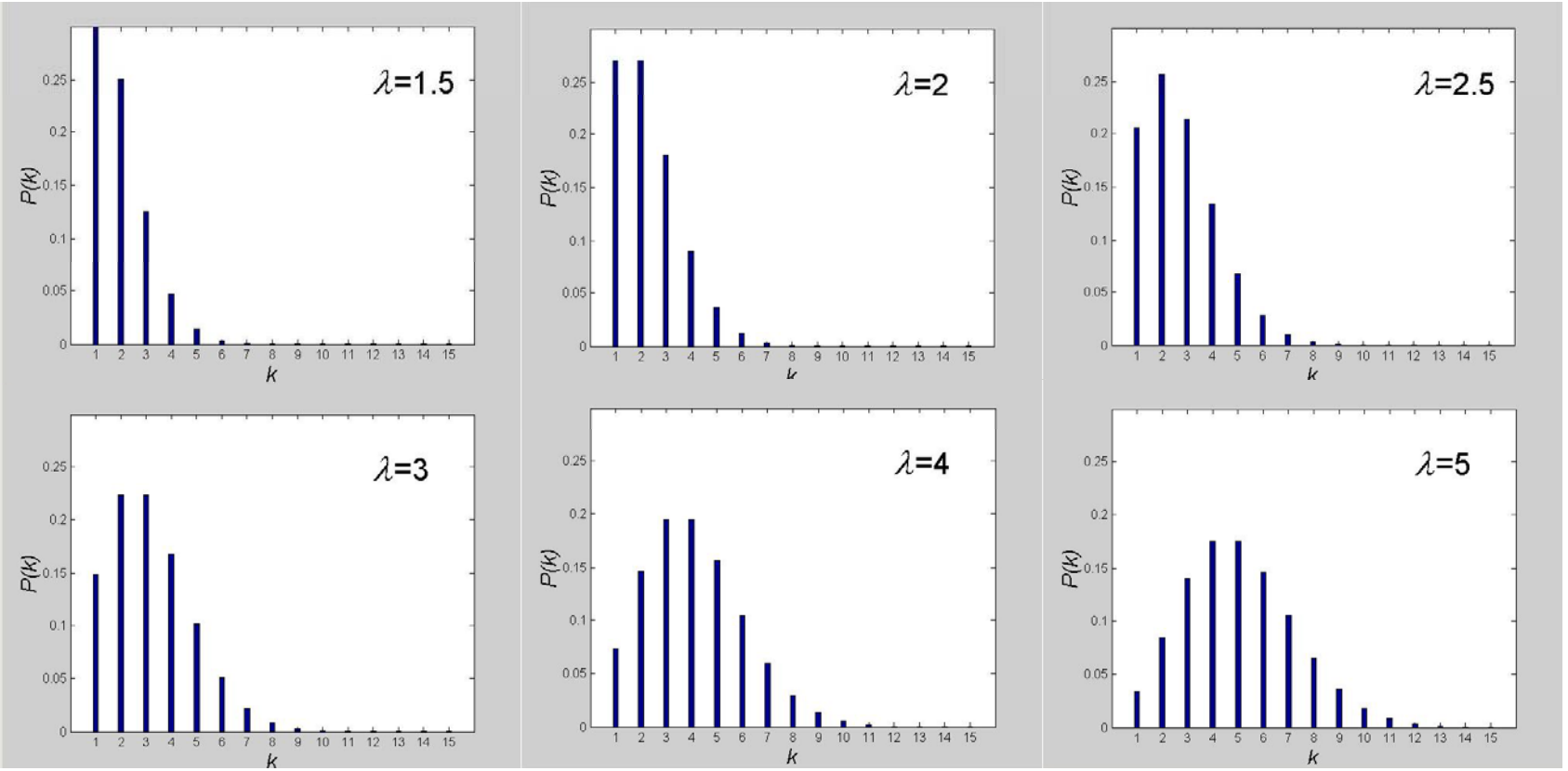
Degree Distribution (DD)



- *P(k)* - probability that a randomly selected node has exactly *k* links
- Most nodes have the number of links close to the average
- Some nodes have fewer links
- Few nodes have significantly larger number of links

K – number of links

Degree Distribution for Random Graphs



k number of links, *P(k)* – pdf, parameter λ is the average degree

- Most nodes have the number of links close to the average
- Few nodes have significantly larger number of links (exponential decay)

Poisson and Power Law DD

- A classical random graph asymptotically has Poisson DD (Degree Distribution)

  here, $P(k) = \frac{e^{-\lambda}\lambda^k}{k!}$ is the average degree

- In our experiments we have found that Power Law $P(k) \sim k^{-\lambda}$ is better approximation

$$\lambda = \overline{k} = \sum_{k=0}^{\infty} k P(k)$$

- Standard procedure to fit experimental data by a Power Law is to use log-log plots

Log-log plots

The random variable x is said to be distributed according to the power-law with the exponent λ if its probability density function satisfies

$$P(x) \sim x^{-\lambda}$$

Power law dependency

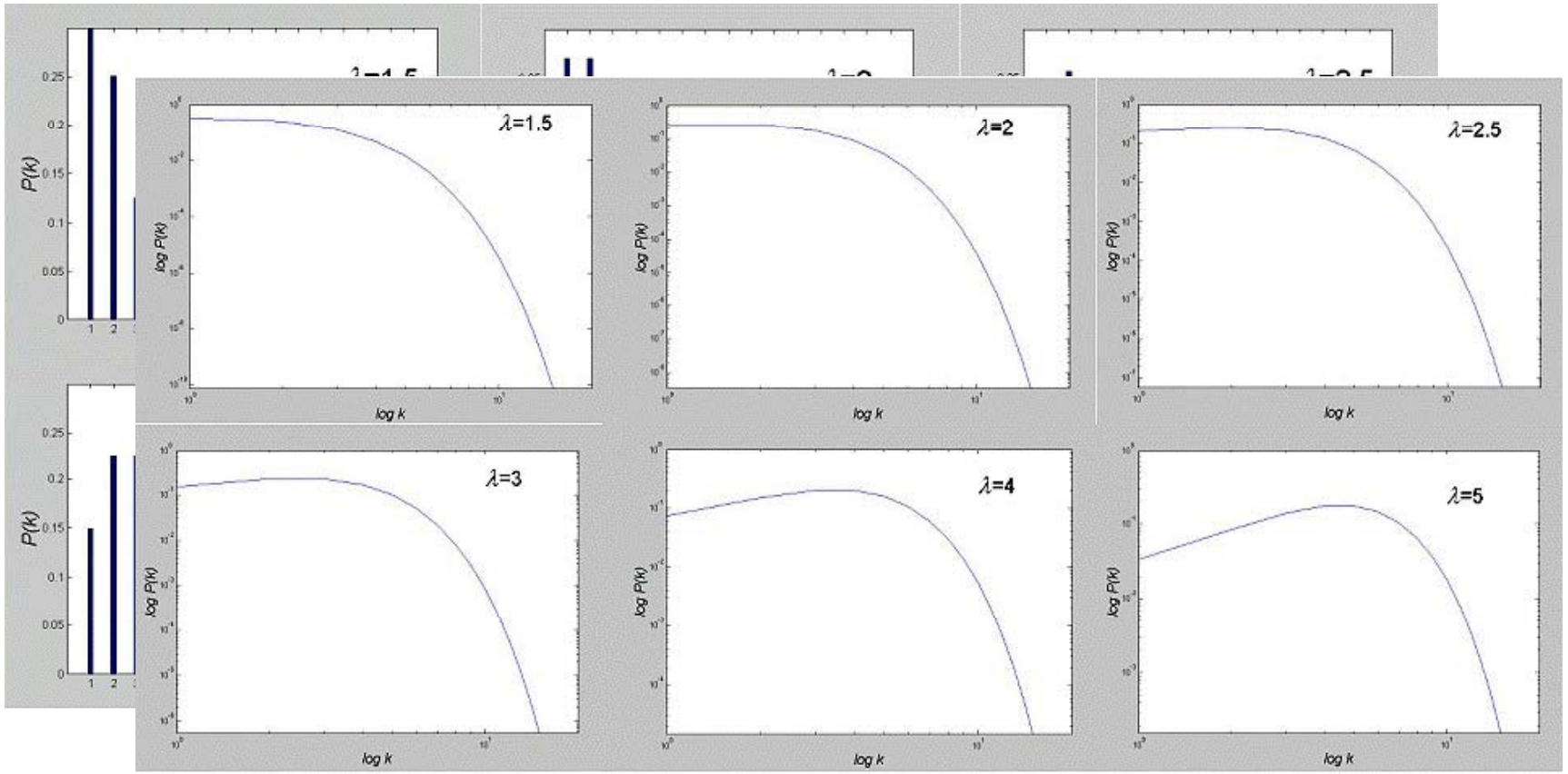
$P(k) \sim k^{-\lambda}$

LOG-LOG

Linear dependency

$\log (P(k)) \sim -\lambda \log (k)$

DD for random graphs: Log-log plots

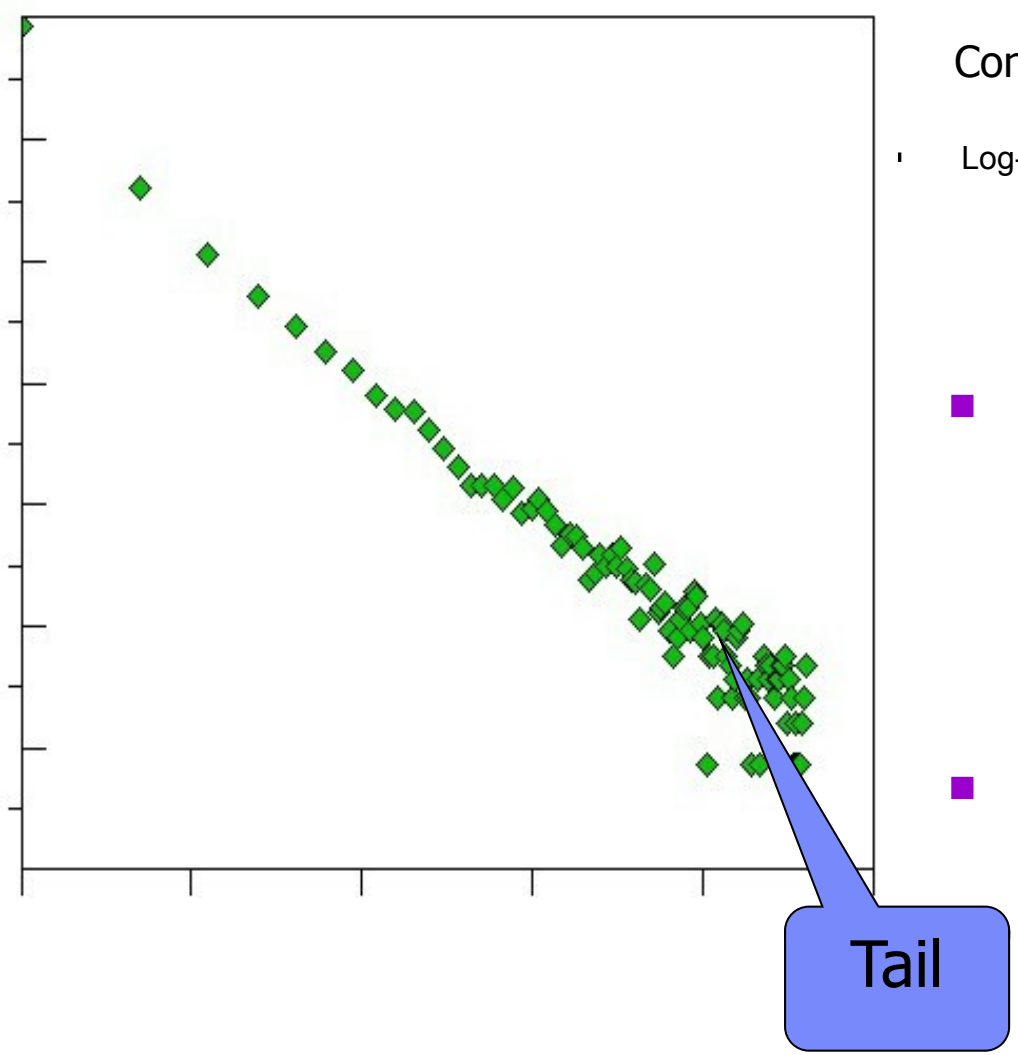


k number of links, *P(k)* – pdf, parameter λ is the average degree

- Unimodal, no linear segments on the Log-Log plot as it would be for the power-law DD: $P(k) \sim k^{-\lambda}$

How to measure Zipf like distributions - Problem

Troussov and O'Donovan "Statistics of Morphological Finite-State Transition Networks Obey the Power Law". RANLP - 2003



Consider In-degree for English dictionary

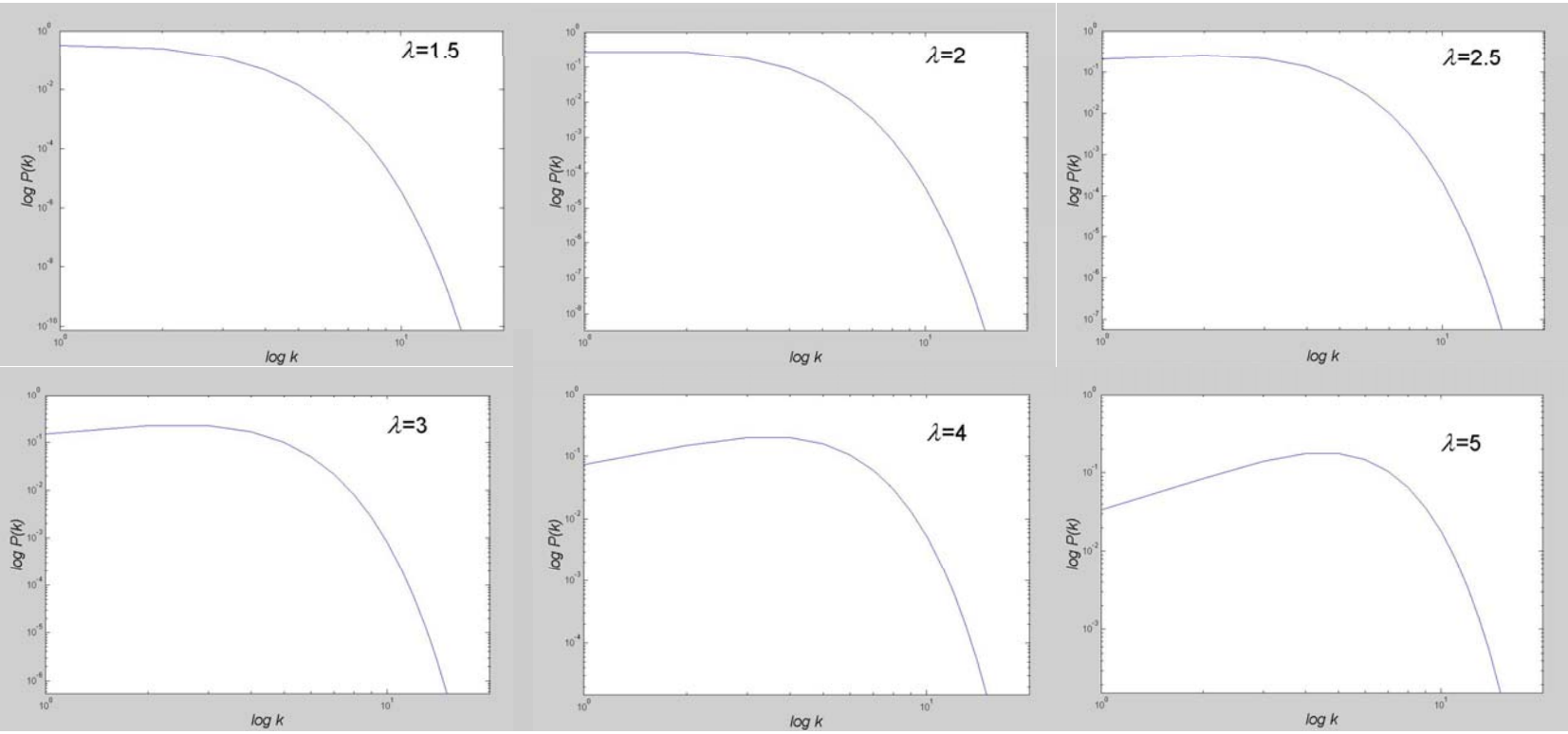
- Log-Log is not enough…

- Not monotonic, very messy at the tail end of the distribution
  - 3 nodes have exactly 103 inc. links, but no nodes with 102
- In *log* scale – few data points at the beginning, too many at the tail

How to measure Zipf like distributions - Solution

- Cumulative distribution
  - Fluctuations become nearly invisible
- To bin the data into exponentially wider bins (they will appear evenly spaced on a log scale)
  - i.e. to divide the region of log k for fitting to equal-size windows and to replace all the points in each window by a single point.

DD for random graphs: Log-log plots



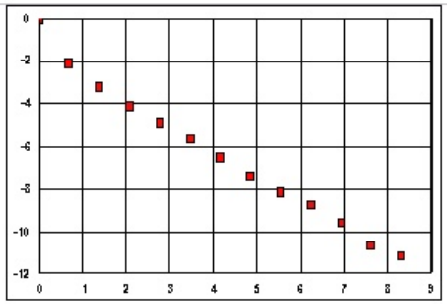
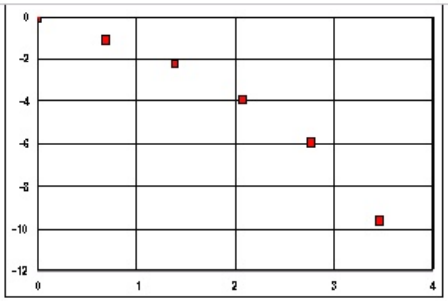
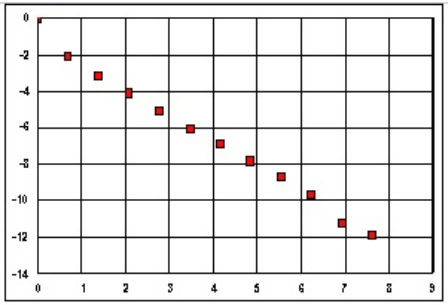
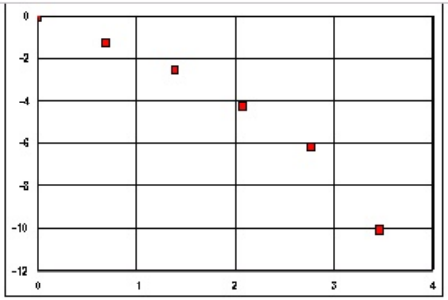
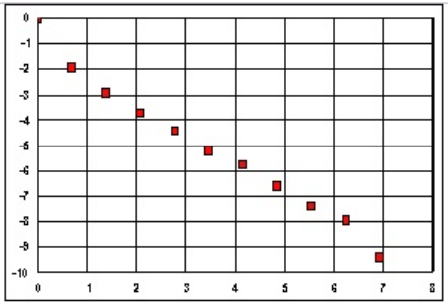
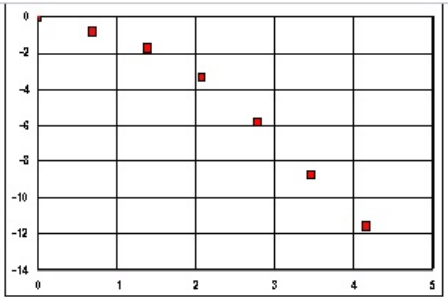
k number of links, *P(k)* – pdf, parameter λ is the average degree

- Unimodal, no linear segments as it would be for the power-law DD: $P(k) \sim k^{-\lambda}$
- Power-law – fat tailed distribution
- Other degree distribution: exponential, multifractal

Poisson DD vs. Power Law DD 2

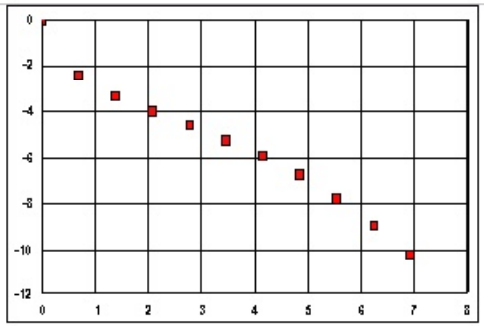
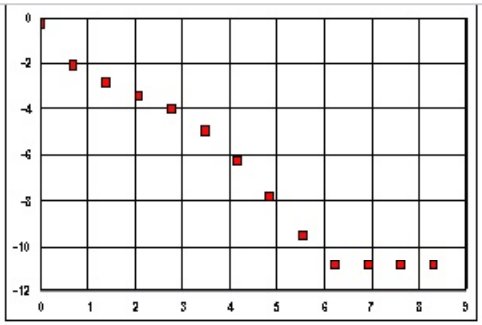
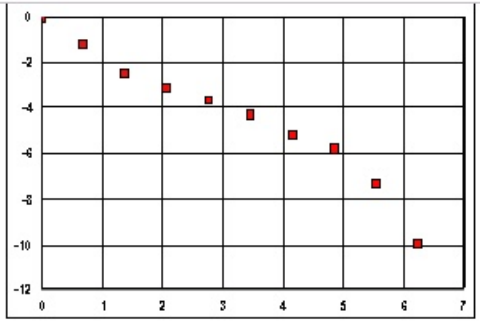
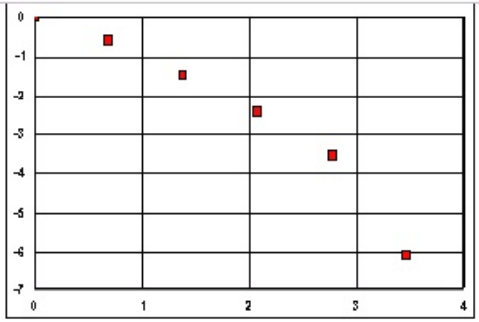
Poisson (Bell Curve) DD	Power Law DD
▪ Example: US National Highway Network	▪ Example: Air traffic system
▪ Peaked distribution.	▪ Continuously decreasing curve
▪ Bell like curves have an exponentially decaying tail; no hubs ▪ Some are “small-world”	▪ Fat-tailed. Power laws decay far more slowly, allowing for “rare events” such as the hubs. ▪ Small-world network (any two airports by 1 to five links)
▪ Most nodes have the same number of links. ME is useful quantity. ▪ Single scale networks	▪ Many small events coexist with a few large events, scale-free; ME diverges ▪ When followed by exp cut-of broad scale networks
▪ Ogranizational principles: several ▪ Constrains limiting the addition of new links	▪ Ogranizational principles: only one (profit) ▪ Preferential linking …

Samples of In- and Out-Degree Distribution

	In-degree	*Out-degree*
English. FSA.		
German. FSA.		
Spanish. FSA.		

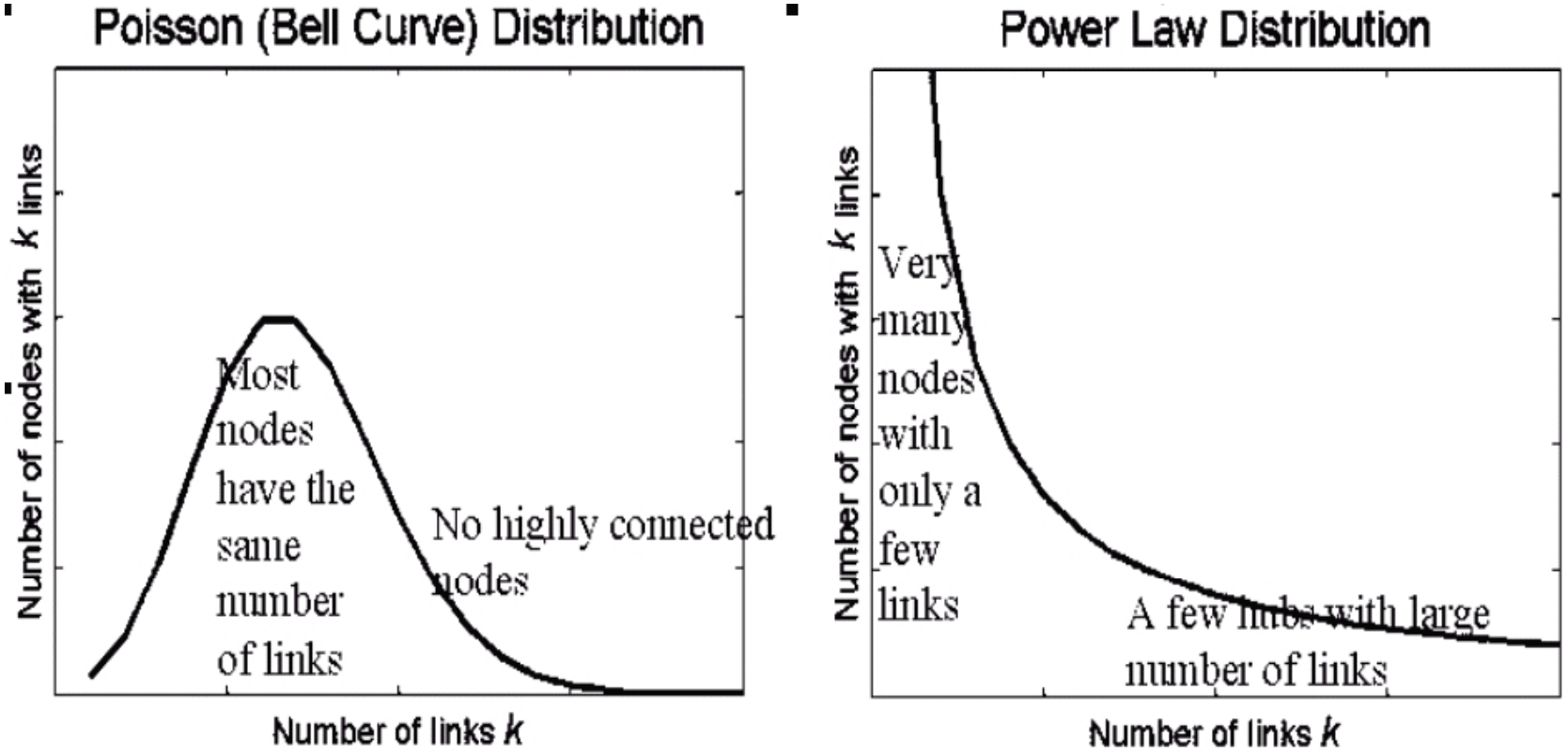
Figurese: Troussov et al. "Per-Node Optimization of Finite-State Mechanisms for Natural Language Processing", Text Processing, Lecture Notes in Computer Science, Volume 2588/2008

Samples of In- and Out-Degree Distribution (Cont.)

	In-degree	*Out-degree*
Chinese traditional. Ideographic language, FSA.		
Finnish. Agglutinative language, finite-state transducer.		

5 sample languages (out of 16 examined ones): 4 languages, whose morphology is represented by FSAs (English, Chinese as an ideographic language, Germanic language and Romance language) and Finnish – an agglutinative language, whose morphology is represented by a lexical transducer implementing two-level morphology.

Shown are log-log (natural log) plots of the cumulative distribution function for in- and out-degree distributions, with data binned into exponentially expanding bins



In Poisson distribution most of the nodes have number of links close the average. In power-law distribution small number of nodes have significant number of links, coexisting with the many nodes which have only a few links.

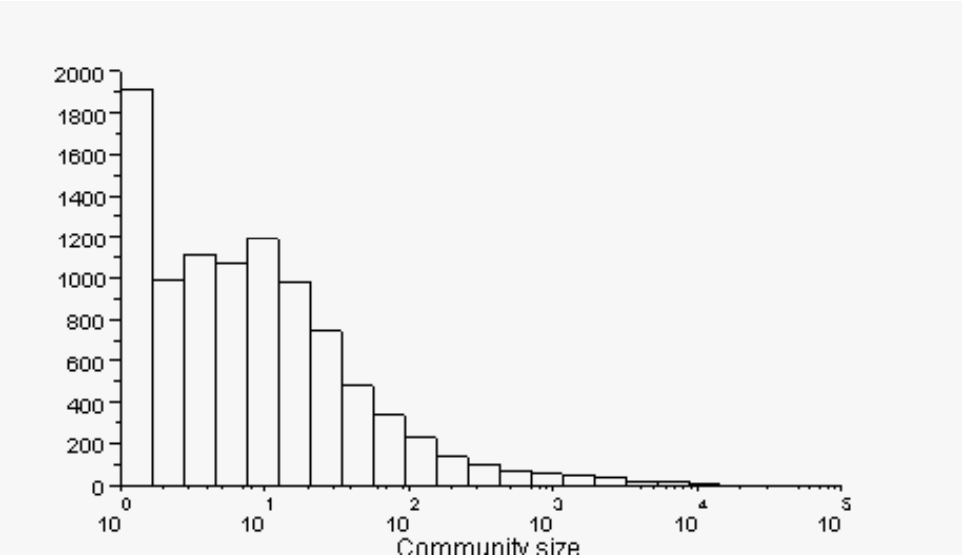
The exponents of power-law approximation

Language	Finite-state device	In-Degree exponent	Out-Degree exponent
Chinese simplified	FSA	2.47	2.53
Chinese traditional	FSA	2.55	2.56
Czech	FST	2.44	2.69
Danish	FSA	2.62	3.19
Dutch	FSA	2.71	3.35
English	FSA	2.52	3.18
Finnish	FST	2.37	2.35
French	FSA	2.50	3.38
German	FSA	2.67	3.32
Italian	FSA	2.43	3.14
Norwegian	FSA	2.56	3.17
Polish	FST	2.56	2.96
Portuguese	FSA	2.42	3.20
Spanish	FSA	2.45	3.20
Swedish	FSA	2.58	3.09
Thai	FST	2.82	2.95

The exponents of power-law approximation of the distribution of in- and out-degree found for our experimental data.

An example of Power Law like Distribution
What is an average Big Blue community?

- Communities formed in IBM in Communities application of IBM Lotus Connections suite:
  - More than 9500 Communities,
    The oldest – 3 years,
    The biggest – 23,000 members
- Distribution of community size is highly skewed
  - with power-law behavior in the upper tail
  - 90% have less than 65 members
    75% have less than 20 members



- Small communities (less than 10 members) are not popular.

Source: Nevidomsky, A. and Troussov, A. "Structure and Dynamics of Enterprise 2.0 Communities", Web Science Conference 2010