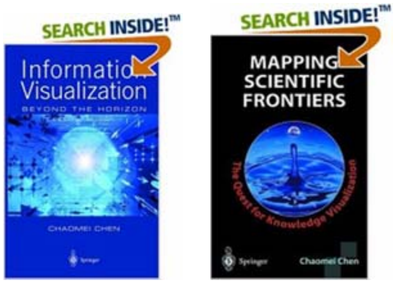
An Explanatory and Computational Theory of Scientific Discovery
可解释性和可计算性的科学发现理论

Chaomei Chen
College of Information Science and Technology, Drexel University
WISELAB, Dalian University of Technology

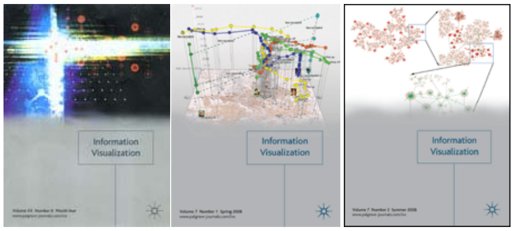
ISTIC, Beijing. Sept 11, 2009

Background Readings

Chen, C. (2004) *Information Visualization: Beyond the Horizon*. Springer. 2nd ed. ISBN: 1-85233-789-3.
Chen, C. (2003) *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. Springer. ISBN: 1-85233-494-0.


SEARCH INSIDE!™
Information Visualization
BEYOND THE HORIZON
CHAOMEI CHEN
Springer
SEARCH INSIDE!™
MAPPING SCIENTIFIC FRONTIERS
The Quest for Knowledge Visualization
Springer
Chaomei Chen

Information Visualization (IVS)


Information Visualization
Information Visualization
Information Visualization

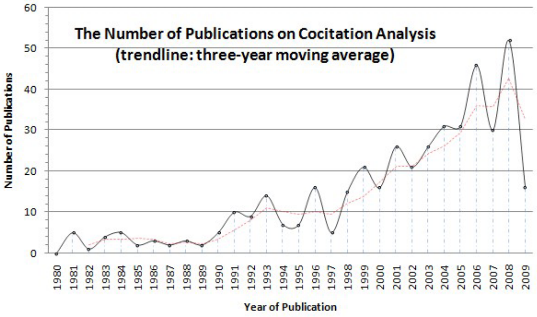
Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3(3), 191-209.

Outline

- 1. Introduction
  - – Motivation of the work
    - Three grand challenges
- 2. The nature of insight
  - – A recurring theme
  - – A mechanism of discovery
- 3. An explanatory theory
  - – Principles of the theory
  - – Examples
    - Scientific discovery
    - Complex network analysis
  - – Implications on knowledge diffusion theories
    - Re-thinking information foraging theory
- 4. Conclusions
  - – Directions and challenges
  - – Conclusions

1. Motivations

- **Quantitative studies of science are increasingly popular …**



Multiple Factors Cause the Increase

- Scientometrics, bibliometrics, informetrics, …
- Web of Science, Scopus, Google Scholar, …
- H-index, G-index, …
- Pajek, UCINet, ManyEyes, …
- HistCite, CiteSpace, …

Grand Challenges

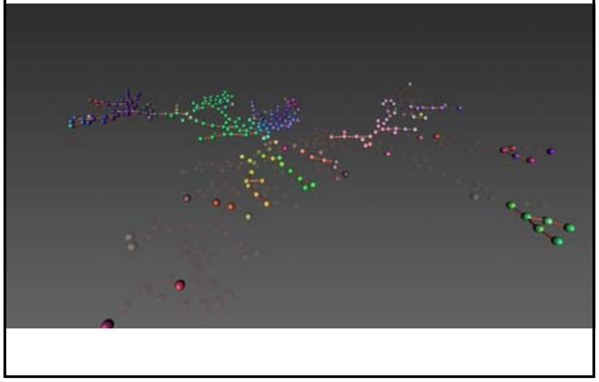
1. Understand emerging trends and essential structures of complex and evolving scientific disciplines, fields, and specialties
   - Theories of scientific change
   - Theories of discovery and innovation
   - Theories of knowledge diffusion
2. Make enabling techniques accessible to everyone, including analysts, scholars, scientists, policy makers, and the public such that they can routinely and repeatedly monitor the development of science
   - Network analysis and visualization
   - Text mining
   - CiteSpace

- Tightly coupled studies of science and practices of science
  - Literature-based discovery
  - Cyber-enabled discovery
  - SDSS

G1

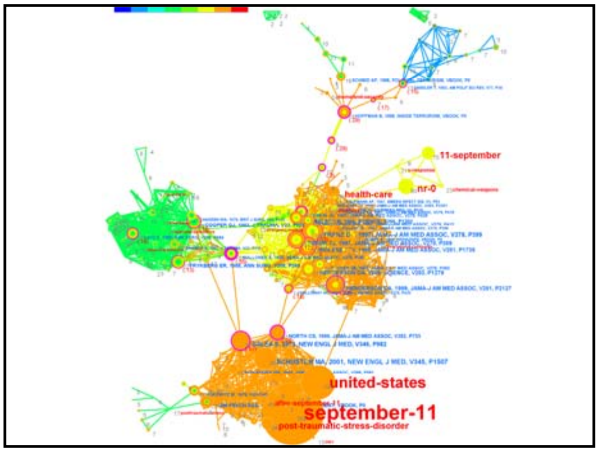
- Global trends and patterns

The growth of superstring research

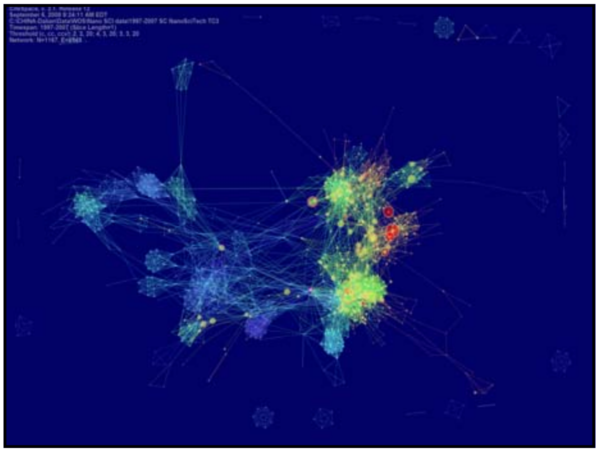


Some animations …

- Superstring research
- Botox research
- BSE
- BSE (transparent)



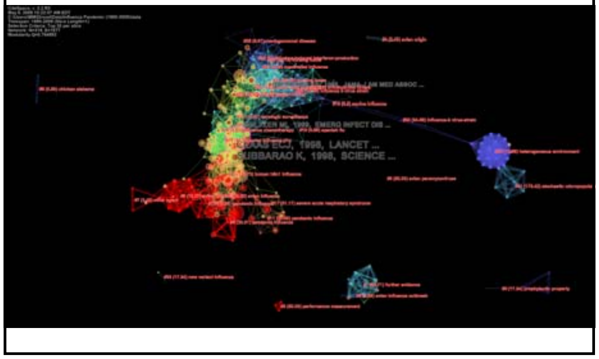
11-september
nr-0
chemical-weapons
health-care
HOFFMAN B, 1998, INSIDE TERRORISM, VBOOK, P0
FRANZ DR, 1997, JAMA-J AM MED ASSOC, V278, P399
HENDERSON DA, 1999, JAMA-J AM MED ASSOC, V281, P2127
NORTH CS, 1999, JAMA-J AM MED ASSOC, V282, P755
GALEA S, 2002, NEW ENGL J MED, V346, P982
SCHUSTER MA, 2001, NEW ENGL J MED, V345, P1507
united-states
after-september-11
september-11
post-traumatic-stress-disorder



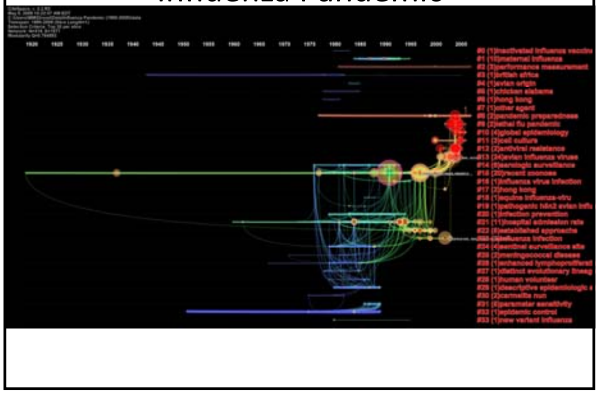
G2

- Easy access to tools

Influenza Pandemic

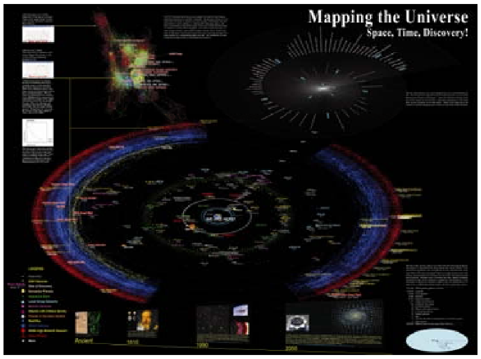

SOUZA ELC, 1998, LANCET ...
SUBBARAO K, 1998, SCIENCE ...
MATSUZAKI Y, 1999, EMERG INFECT DIS ...

Influenza Pandemic



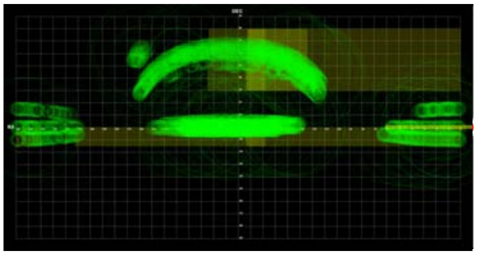
G3

- Tightly coupled studies of science and scientific research



The *Mapping the Universe* won 2008 NSF International Science & Engineering Visualization Challenge Semi-finalist.
The map is included in *The Places & Spaces* Exhibit, traveling more than ten countries and 30 cities.

Accessing Scientific Data



A 2-D projection of the Universe with the spatial pattern of SDSS query log

1. Summary

- 3 grand challenges that motivate the work in longer terms
- These challenges highlight the need of
  - Theories of scientific change
  - Theories of scientific discovery
  - Theories of knowledge diffusion
  - Tools for new ways to explore and interact with scientific knowledge
  - Tools for seamlessly working with both scientific data and bibliographic data

2. The Nature of Insight and Creativity

- *"Creativity is the friction of the attention space at the moments when the structural blocks are grinding against one another the hardest"*

  Collins 1998, p. 76

- The philosophers of greatest repute tended to be rivals representing ***conflicting*** schools of thought for their generation.

What We Know about Scientific Change

- 1962
  - New paradigms are typically initiated by ***young*** scientists or ***newcomers*** to the crisis-laden field (Kuhn, 1962). What do they have in common?
- 1969
  - What can scientists do to keep up their creativity?
  - Scientists maintain contacts with scientists and scientific work in areas ***different from their own*** in order to enhance their ability to develop new ideas in their own areas (Crane, 1969).
- 1977
  - A good system for knowledge creation may not be good for knowledge diffusion; and vice versa.
  - Major changes in a variety of disciplines tend to be generated within small, socially ***coherent*** groups (Griffith & Mullins, 1977).
- 2007
  - Brokerage leads to greater collaborative creativity
    - Fleming, Mingo, & Chen, 2007 tested in a study of collaborative inventors of utility patents
  - Cohesive networks hamper creativity but aid in its transfer, particularly if the knowledge is complex and tacit.
  - New combinations, as integrative work, are defined as a mechanism of creativity.
- 2009
  - Our theory focuses on transformative discoveries
    - conceptually more complex than new combinations of existing discoveries.
    - new concepts and theories must be introduced before integrative work becomes possible.

Knowledge diffusion models

- Epidemic models
  - The contact rate between scientists can speed up the diffusion of knowledge.
- Ant colony models
  - Dorigo & Gambardella, 1997
  - Ants travel between their home and food sources.
  - They leave scents as trails for others.
  - Scents decrease over time unless being reinforced by other ants.
  - Ants → scientists
  - Their home → the contemporary intellectual structure.
  - The food sources → new publications in the literature.
  - Finding foods → making a reference to a new publication and leaving trails for other scientists.
- Random walk models
  - Each node in the network represents a state.
  - Moving from one node to another is governed by a state transition probability.
  - The spread of knowledge is thus translated into a question of how easy or how hard one could make such moves.
- Information foraging
  - Traditionally it has not been seen as a knowledge diffusion model.
  - However, our new theory of discovery provides a broader framework in which one can turn the information foraging theory into a knowledge diffusion model.
- Most knowledge creation theories and knowledge diffusion theories are separated
  - If a single theory could explain both, that would be favorable.
  - Our theory provides new interpretations of these diffusion models.

Predicting Nobel prizes:

Quantifying a Nobel-prize worthy research

- Citation Counts:
  - Between 1977 and 1992, Garfield studied Nobel prize winners' publications and their citations.
    - among 100 most cited authors from 1981 through 1990, eight Nobel laureates appeared on the list (Garfield & Welljamsdorof, 1992).
    - Others on the list are regarded as potential future Nobel prize winners.
- h-index
  - Another simple metric to quantify the impact of individual scientists.
- IQp - (Antonakis & Lalive, 2008).
  - An index of the quality and productivity of a scholar
    - The number of citations
    - The number of papers
    - Academic age
    - Top 3 subject categories in which one has been most cited
  - Tested the new index on Nobel winners in physics, chemistry, medicine, and economics.
    - IQp =5 => tenure
    - IQp > 20 => very important influence on the field
    - IQp > 60 => Nobel prize
- Strengths
  - Simplicity
- Weaknesses
  - They do not present deeper insights into the nature of scientific discovery.

Literature-based discovery

- Don Swanson
  - identify potentially valuable hypotheses (Swanson, 1986a, 1986b; Swanson, 1987; Swanson & Smalheiser, 1999)
  - discovery from public knowledge
    - Given A~B and B~C, is A~C a reasonable hypothesis?
      - fish oil and Raynaud's syndrome (Swanson, 1986a)
        - » *blood viscosity* bridges *Raynaud's disease and dietary fish oil.*
      - magnesium and migraine (Swanson, 1988)
      - indomethacin and Alzheimer's disease (Smalheiser & Swanson, 1996).
- Further Development
  - Gordon & Lindsay, 1996; Lindsay & Gordon, 1999
  - Using lexical statistics to discover hidden connections in the medical literature
  - ***Hidden connections are those that are unlikely to be found by examination of bibliographic citations or the use of standard indexing methods and yet establish a relationship between topics that might profitably be explored by scientific research.***

Thinking Outside the Box

- Effective strategies for making scientific discoveries have highlighted the ability to think creatively and ***look at a problem from a fresh perspective***.
- In 1993, Dunbar compared two different strategies of hypothesis generation using a Nobel Prize winning discovery as the test case (Dunbar, 1993).
  - It is a more effective discovery strategy to encourage researchers to consider novel alternative hypotheses.
- In 2007, Heinze & Bauer did a longitudinal study of highly creative scientists in nano science and technology
  - It is not only the sheer quantity of publications that enables scientists to produce creative work but also their ability to ***effectively communicate with otherwise disconnected peers and to address a broader work spectrum*** (Heinze & Bauer, 2007).

2. Summary

- Many discoveries establish a new connection between bodies of knowledge
  - E.g. Literature-based discovery
- Many good discoveries are quickly recognized and rapidly spread across scientific communities.
  - There are also many exceptions

3. An Explanatory Theory

- Key principles
  - Structural properties
    - Brokerage as a discovery mechanism
  - Temporal properties
    - Good ideas are easy to recognize

Questions

- Why is it possible that communicating with otherwise disconnected scientists can lead to more creative work?
- What can one do specifically to come up with novel alternative hypotheses?
- How do we think outside the box?

Structural Holes

- The Concept
  - The presence of a structure hole in a social network is due to ***sparse connections*** among individual nodes (Burt, 1992, 2001, 2004).
- The Theory
  - The information flow in the network is limited by the topology of the network.
  - Some positions in the network are more privileged and advantageous than others in terms of access to information.
  - People at such privileged positions, or gatekeepers, inherit advantages of their positions.
- Evidence
  - Burt has shown that ***creative ideas are more likely to appear*** at such gatekeepers' positions than elsewhere in a network.

Structural Holes in Social Networks

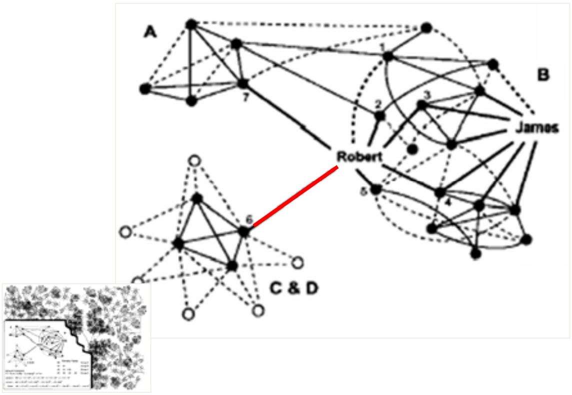
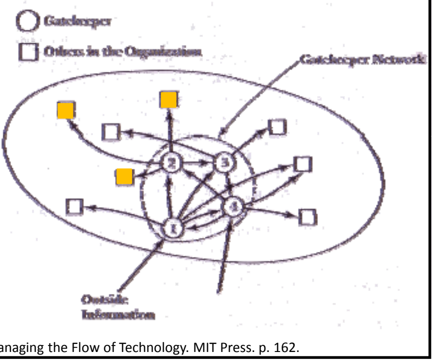
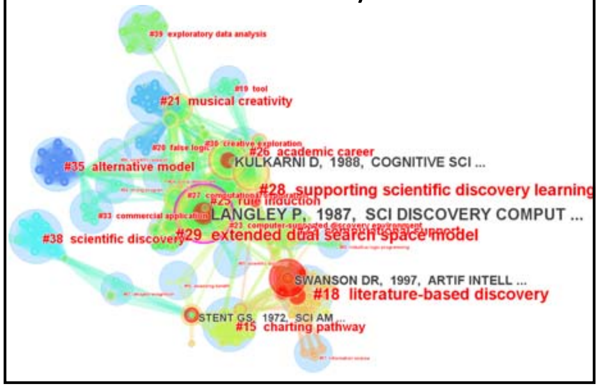


Figure 6.14 The Functioning of the Gatekeeper Network. New information is brought into the organization by 1. It can be transmitted to 2, 3, and 4 via the gatekeeper network. It reaches its eventual users (indicated by squares) through their contacts with gatekeepers.

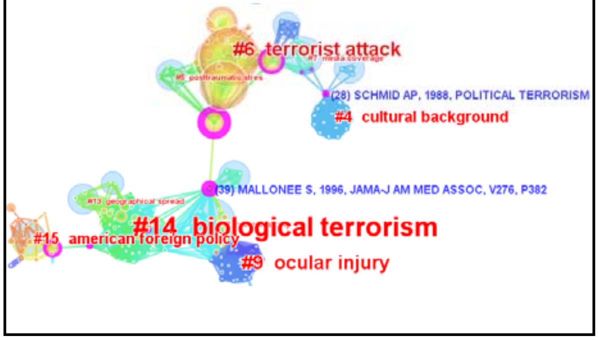


Thomas J. Allen (1977) Managing the Flow of Technology. MIT Press. p. 162.

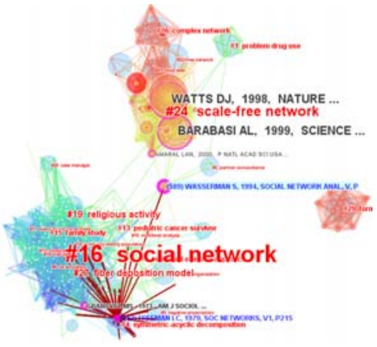
Scientific Discovery Research


#39 exploratory data analysis
#19 tool
#21 musical creativity
#20 false logic
#30 creative exploration
#26 academic career
#35 alternative model
KULKARNI D, 1988, COGNITIVE SCI ...
#28 supporting scientific discovery learning
#27 computational exploration
#25 rule induction
#33 commercial application
LANGLEY P, 1987, SCI DISCOVERY COMPUT ...
#23 computer-supported discovery environment
#38 scientific discovery
#29 extended dual search space model
SWANSON DR, 1997, ARTIF INTELL ...
#18 literature-based discovery
STENT GS, 1972, SCI AM ...
#15 charting pathway

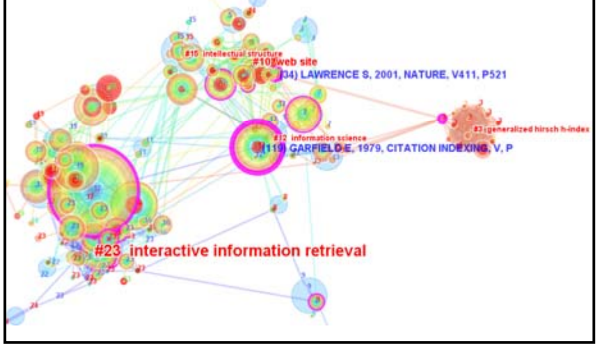
Terrorism Research



Social network and complex network

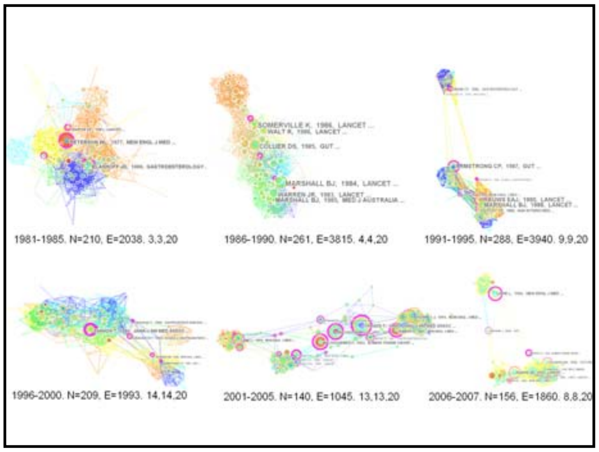

complex network
#1 problem drug use
WATTS DJ, 1998, NATURE ...
#24 scale-free network
BARABASI AL, 1999, SCIENCE ...
AMARAL LAN, 2000, P NATL ACAD SCI USA ...
(589) WASSERMAN S, 1994, SOCIAL NETWORK ANAL, V, P
#19 religious activity
#15 family study
#13 pediatric cancer survivor
#16 social network
#20 fiber deposition model
AM J SOCIOL ...
FREEMAN LC, 1979, SOC NETWORKS, V1, P215
symmetric acyclic decomposition

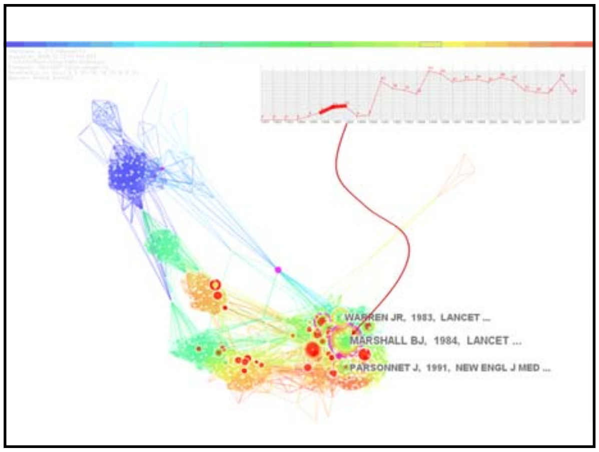
Information Science


#15 intellectual structure
#10 web site
(34) LAWRENCE S, 2001, NATURE, V411, P521
#3 generalized hirsch h-index
#12 information science
(119) GARFIELD E, 1979, CITATION INDEXING, V, P
#23 interactive information retrieval

Case Study 1: Peptic Ulcer

- The Nobel Prize in Physiology or Medicine for 2005 was awarded jointly to Barry J. Marshall and J. Robin Warren for their discovery of "the bacterium *Helicobacter pylori and its role in* gastritis and peptic ulcer disease."


SOMERVILLE K, 1986, LANCET ...
WALT R, 1986, LANCET ...
COLLIER DS, 1985, GUT ...
MARSHALL BJ, 1984, LANCET ...
WARREN JR, 1983, LANCET ...
MARSHALL BJ, 1985, MED J AUSTRALIA ...
ARMSTRONG CP, 1987, GUT ...
MARSHALL BJ, 1988, LANCET ...
1981-1985. N=210, E=2038. 3,3,20
1986-1990. N=261, E=3815. 4,4,20
1991-1995. N=288, E=3940. 9,9,20
1996-2000. N=209, E=1993. 14,14,20
2001-2005. N=140, E=1045. 13,13,20
2006-2007. N=156, E=1860. 8,8,20

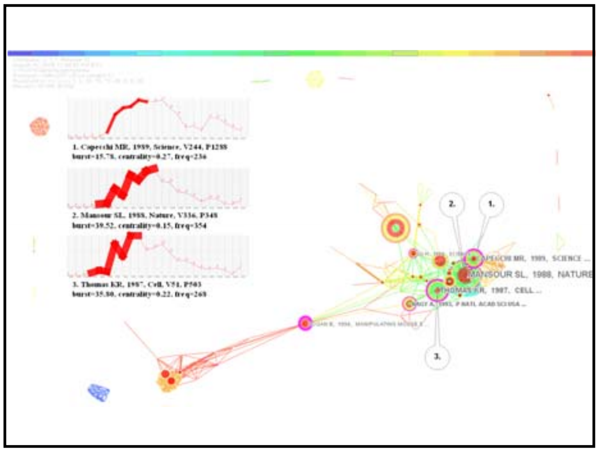

WARREN JR, 1983, LANCET ...
MARSHALL BJ, 1984, LANCET ...
PARSONNET J, 1991, NEW ENGL J MED ...

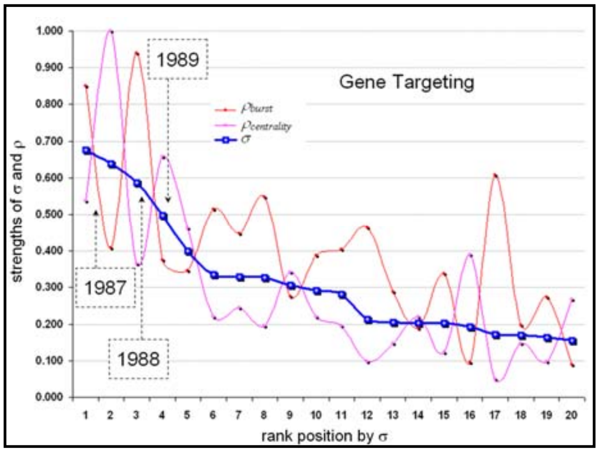
Top 5 most cited references in peptic ulcer research (1980-2007)

Citation	Author	Year	Source	Vol.	Page	ρ_{burst}	$\rho_{centrality}$	σ_2
711	**MARSHALL BJ**	**1984**	**LANCET**	**1**	**1311**	**0.138**	**0.393**	**0.232**
581	PARSONNET J	1991	NEW ENGL J MED	325	1127	0.208	0.143	0.172
579	WARREN JR	1983	LANCET	1	1273	0.165	0.250	0.203
466	YAMADA T	1994	JAMA	272	65	0.635	0.071	0.213
421	MARSHALL BJ	1988	LANCET	2	1437	0.607	0.286	0.416

Case Study 2: Gene Targetting

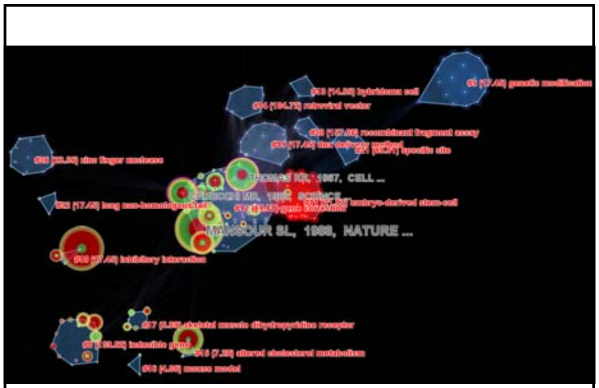
- The Nobel Prize in Physiology or Medicine for 2007 was awarded jointly to Mario R. Capecchi, Martin J. Evans and Oliver Smithies for their discoveries of "*principles for introducing specific gene modifications in mice by the use of embryonic stem cells.*"


1. Capecchi MR, 1989, Science, V244, P1288
burst=15.78, centrality=0.27, freq=236
2. Mansour SL, 1988, Nature, V336, P348
burst=39.52, centrality=0.15, freq=354
3. Thomas KR, 1987, Cell, V51, P503
burst=35.30, centrality=0.22, freq=268
2.
1.
3.
CAPECCHI MR, 1989, SCIENCE ...
MANSOUR SL, 1988, NATURE
THOMAS KR, 1987, CELL ...
P NATL ACAD SCI USA ...
1994, MANIPULATING MOUSE E...

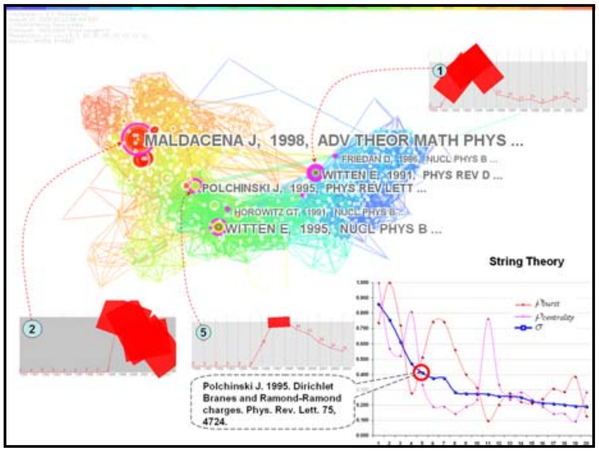


Top 5 references by σ_2 – the geometric mean of centrality and burstness

Author	Year	Source	Vol.	Page	Citations	ρ_{burst}	$\rho_{centrality}$	σ_2
THOMAS KR	1987	CELL	51	503	268	0.851	0.537	**0.676**
HOGAN B	1994	MANIPULATING MOUSE E	BOOK		136	0.409	1.000	**0.639**
MANSOUR SL	1988	NATURE	336	348	354	0.940	0.366	**0.586**
CAPECCHI MR	1989	SCIENCE	244	1288	236	0.375	0.659	**0.497**
NAGY A	1993	P NATL ACAD SCI	90	8424	182	0.346	0.463	**0.400**


genetic modification
hybridoma cell
retroviral vector
recombinant fragment assay
specific site
zinc finger nuclease
THOMAS KR, 1987, CELL ...
CAPECCHI MR, 1989, SCIENCE
embryo-derived stem-cell
gene targeting
MANSOUR SL, 1988, NATURE ...
inhibitory interaction
skeletal muscle dihydropyridine receptor
inducible gene
altered cholesterol metabolism
mouse model

Case Study 3: String Theory


1
MALDACENA J, 1998, ADV THEOR MATH PHYS ...
FRIEDAN D, 1986, NUCL PHYS B ...
WITTEN E, 1991, PHYS REV D ...
POLCHINSKI J, 1995, PHYS REV LETT ...
HOROWITZ GT, 1991, NUCL PHYS B ...
WITTEN E, 1995, NUCL PHYS B ...
String Theory
2
5
ρ_{burst}
$\rho_{centrality}$
σ

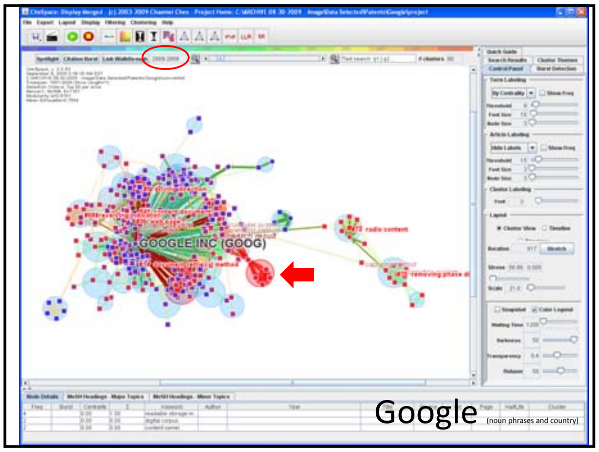
Polchinski J. 1995. Dirichlet Branes and Ramond-Ramond charges. Phys. Rev. Lett. 75, 4724.

The Nature of Maldacena-1998

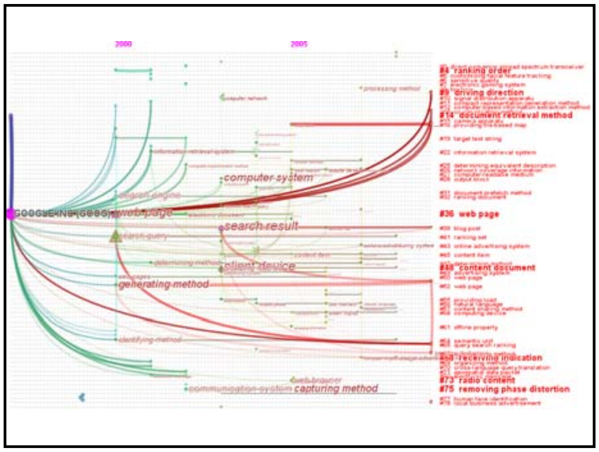
- We aked Juan Maldacena to identify the nature of his major contributions in this article to String Theory.
- His reply: "It *connected* two different kinds of theories: 1) particle theories or gauge theories and 2) string theory. Many of the papers on string dualities (and this is one of them) connect different theories. ***This one connects string theory to more conventional particle theories***."
- TIME 100 Innovator website
  - "he forged a *connection* between the esoteric formulas of string theory and the rest of mainstream physics."
  - "he has been able to suggest a way to knit together ***two theories previously thought to be incompatible***: quantum mechanics, which deals with the universe at its smallest scales; and Einstein's general theory of relativity, which deals with the very largest."
- He is the recipient of the 2007 Dannie Heineman Prize for Mathematical Physics
  - "for profound developments in Mathematical Physics that have illuminated *interconnections* and launched major research areas in Quantum Field Theory, String Theory, and Gravity."

4. Conclusions

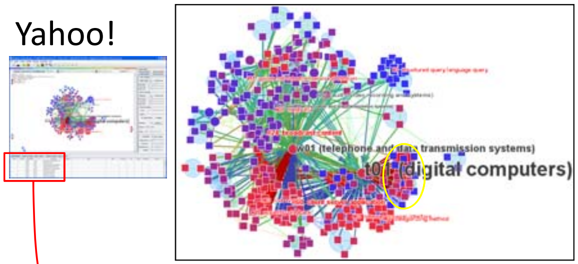
- Three grand challenges
  - Understanding complex and dynamic scientific change
  - Make enabling techniques accessible to everyone
  - Tightly couple studies of science and scientific research
- This is just the beginning!



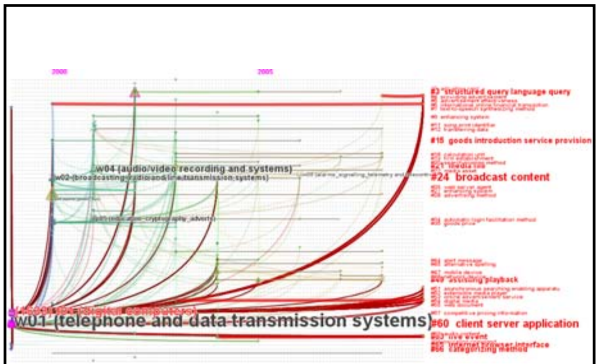
Google (noun phrases and country)



Yahoo!



Freq	Burst	Centrality	Σ	Keyword
8		0.00	1.00	providing advertisement
3		0.00	1.00	priority score
3		0.00	0.00	media player application
3		0.02	1.00	internet advertisement
3		0.00	0.00	extensible media player
2		0.00	0.00	remote computer
2		0.00	0.00	received request
2		0.00	1.00	position auction
2		0.00	0.00	outcome variable



Acknowledgements

NSF Award # IIS-0612129
SEI: Coordinated Visualization and
Analysis of Sky Survey Data and
Astronomical Literature

North-East Visualization
and Analytics Center

中国长江学者计划 (Chang Jiang Scholar Visiting Professor)

大连理工大学，WISELAB, Dalian, China