

# Vision, Resources, and Opportunities for Next Generation of Data Mining and Cyber-Enabled Discovery and Innovation

**Maria Zemankova, Program Director**

[mzemanko@nsf.gov](mailto:mzemanko@nsf.gov)

**Information Integration & Informatics Cluster (IIC)**

**Sylvia Spengler, Cluster Lead**

**Information & Intelligent Information Division (IIS)**

**Haym Hirsh, Division Director**

**Computer & Information Science & Engineering Directorate (CISE)**

**Jeannette Wing, Assistant Director**

**National Science Foundation (NSF)**

**Arden Bement, Director**



National Science Foundation  
WHERE DISCOVERIES BEGIN

# “We”

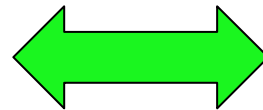
## **You: researchers**

Academia

Industry

Government labs

International



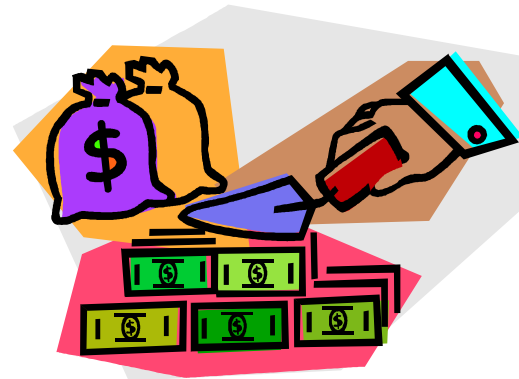
## **Us: supporters**

NSF, NASA, NIH, DOE,...

Industry

Private funders

International



# What “us” need from you

- Research challenges
- Infrastructure needs
- Recommendations for partnerships
- Innovative proposals
  - Research
  - Workshops
  - Educational activities
- Dedication to education, outreach
- Taking broader impacts seriously
- Participation in review process
- Examples of successful research
- Come to NSF as program director, division director,...

# Examples of Successful Research

- Inform the public
- Justify funds spent
- Request more funding

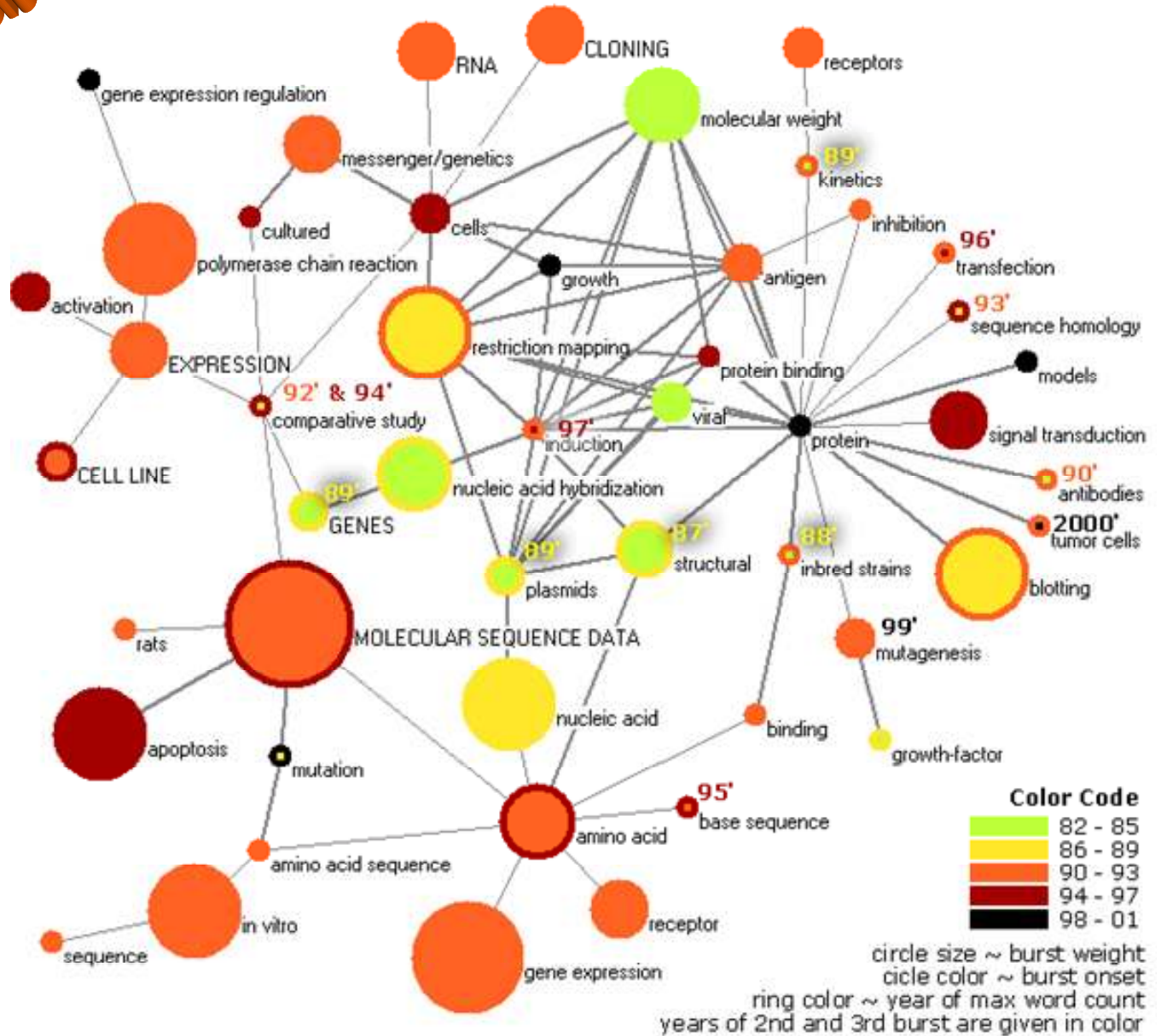
Katy Borner, Indiana University  
 CAREER: Visualizing Knowledge Domains

# Topic Bursts

## Mapping

Visualization of keywords appearing in PNAS 1982 -- 2001.

The top 10% of highly cited PNAS papers were analyzed to reveal keywords listed with high frequency or rapid initial appearance (Topic burst).

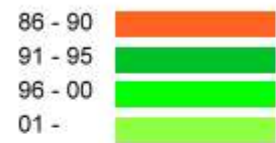


Ke, Visvanath & Bömer  
*Mapping the Evolution of Co-Authorship Networks*

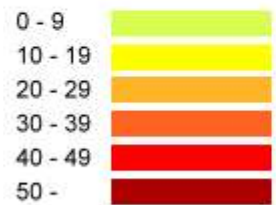
(Won 1st price at the IEEE InfoVis Contest, 2004)

**Color Code:**

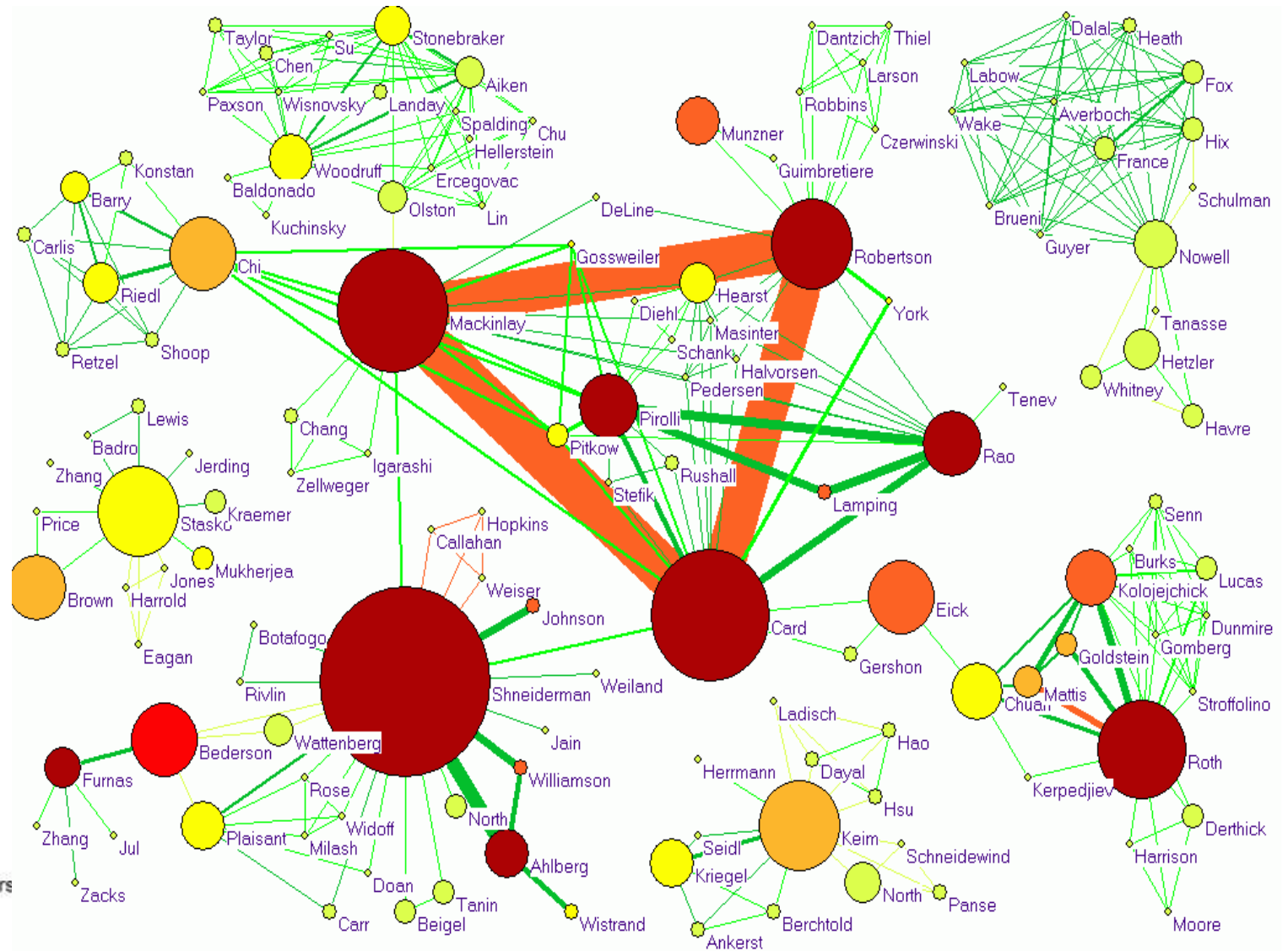
**Line color**



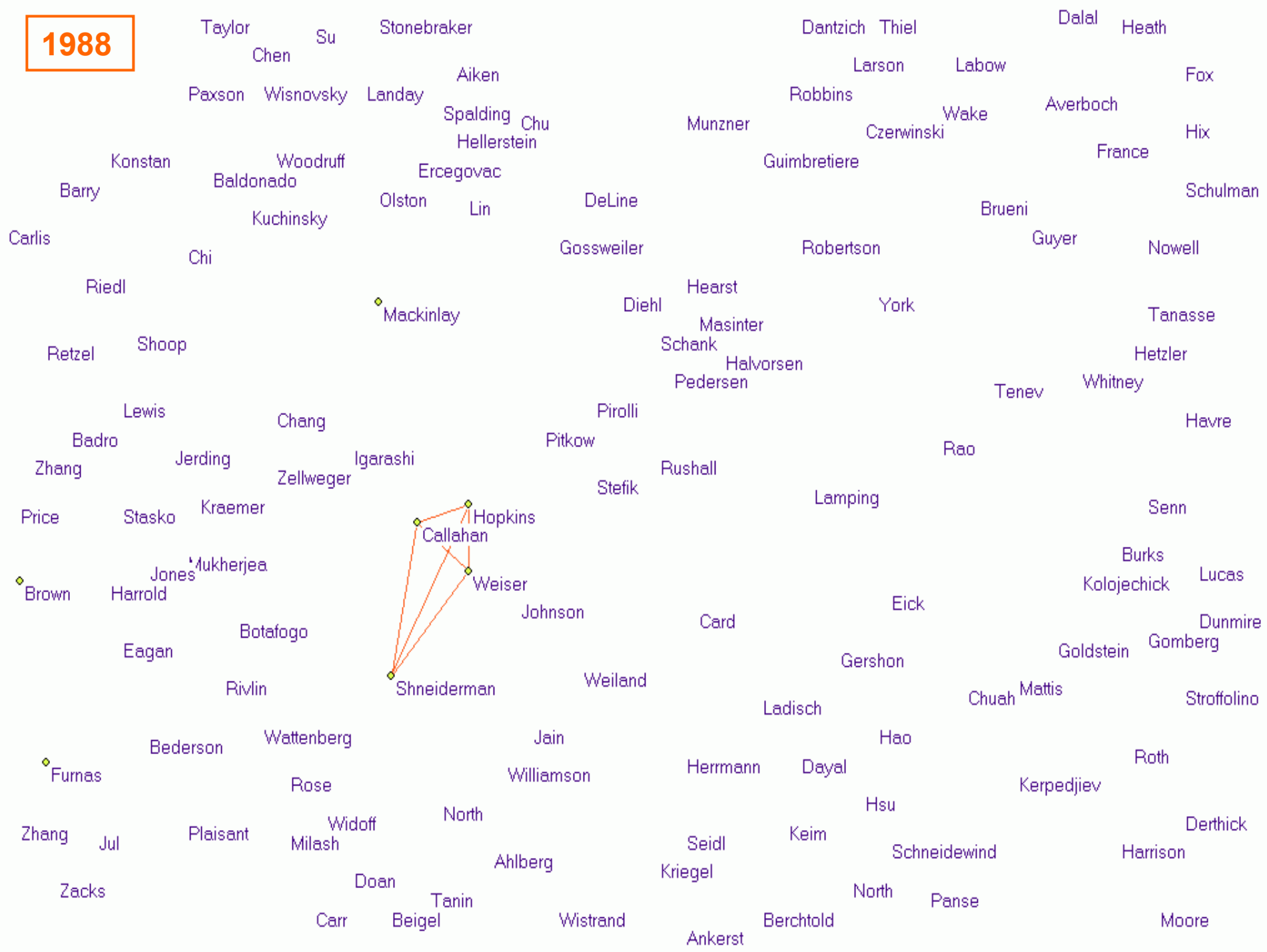
**Node color**



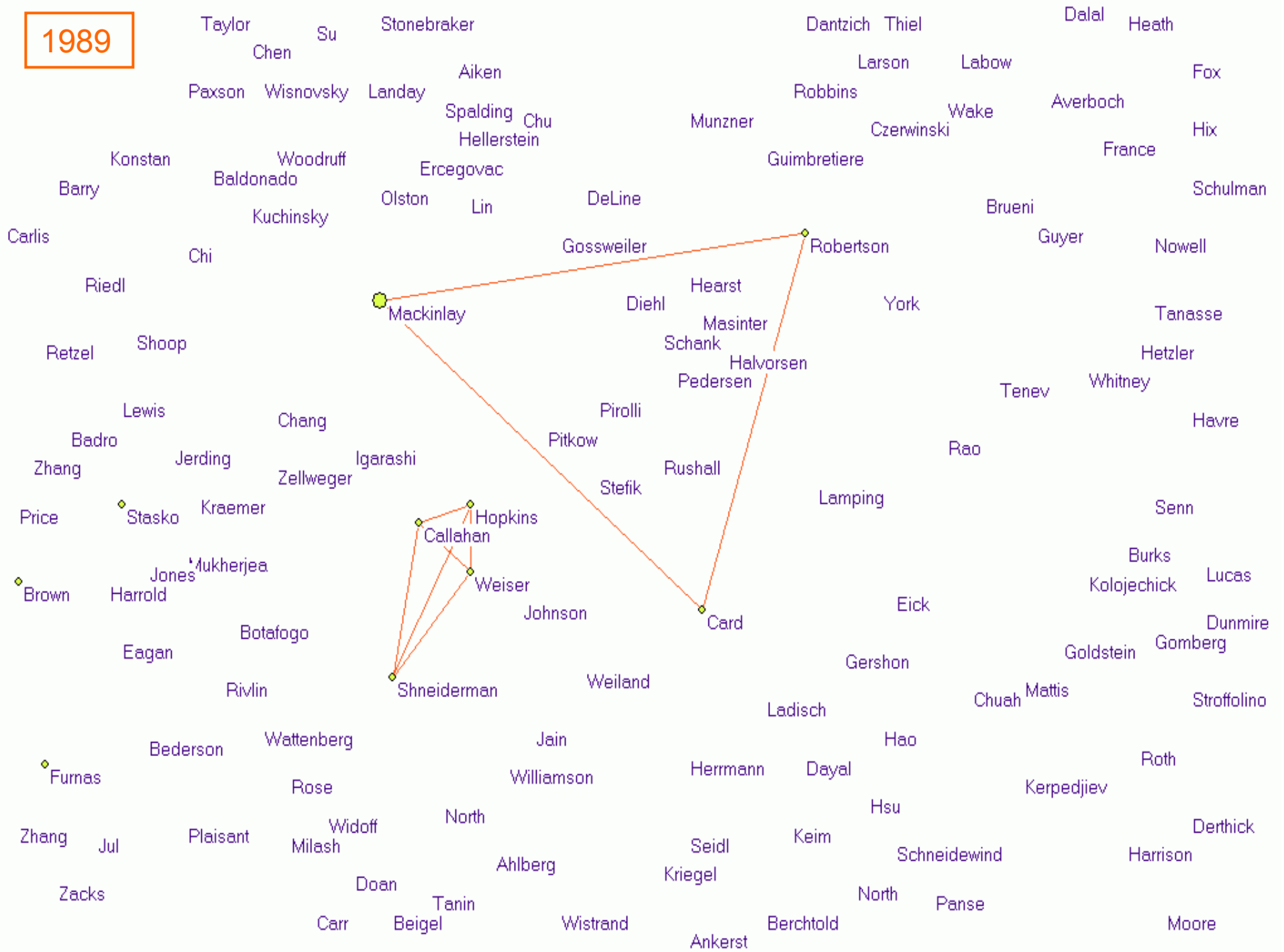
line color ~ year of first time co-authors  
 node color ~ number of citations  
 node size ~ number of papers



1988

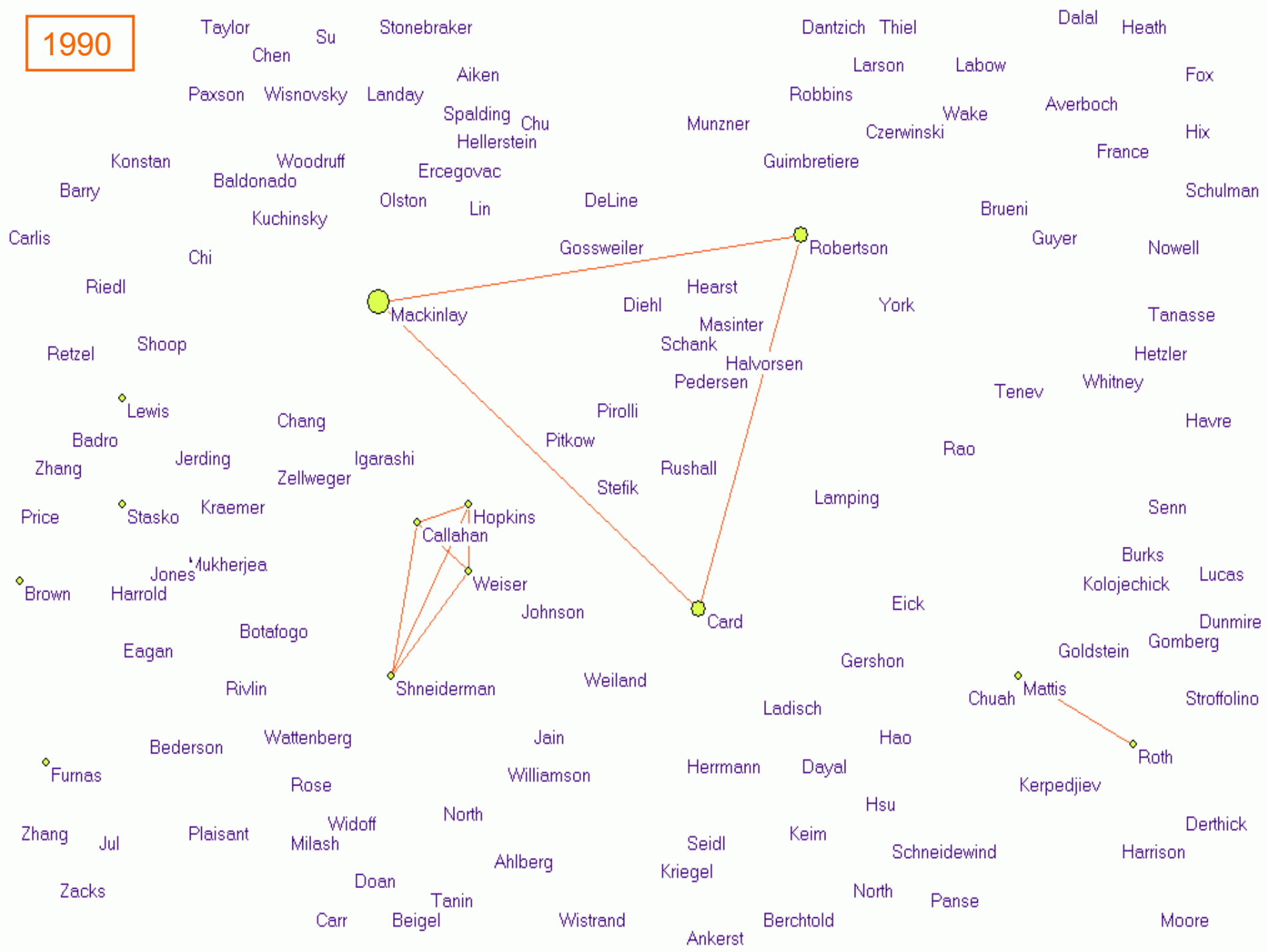


1989

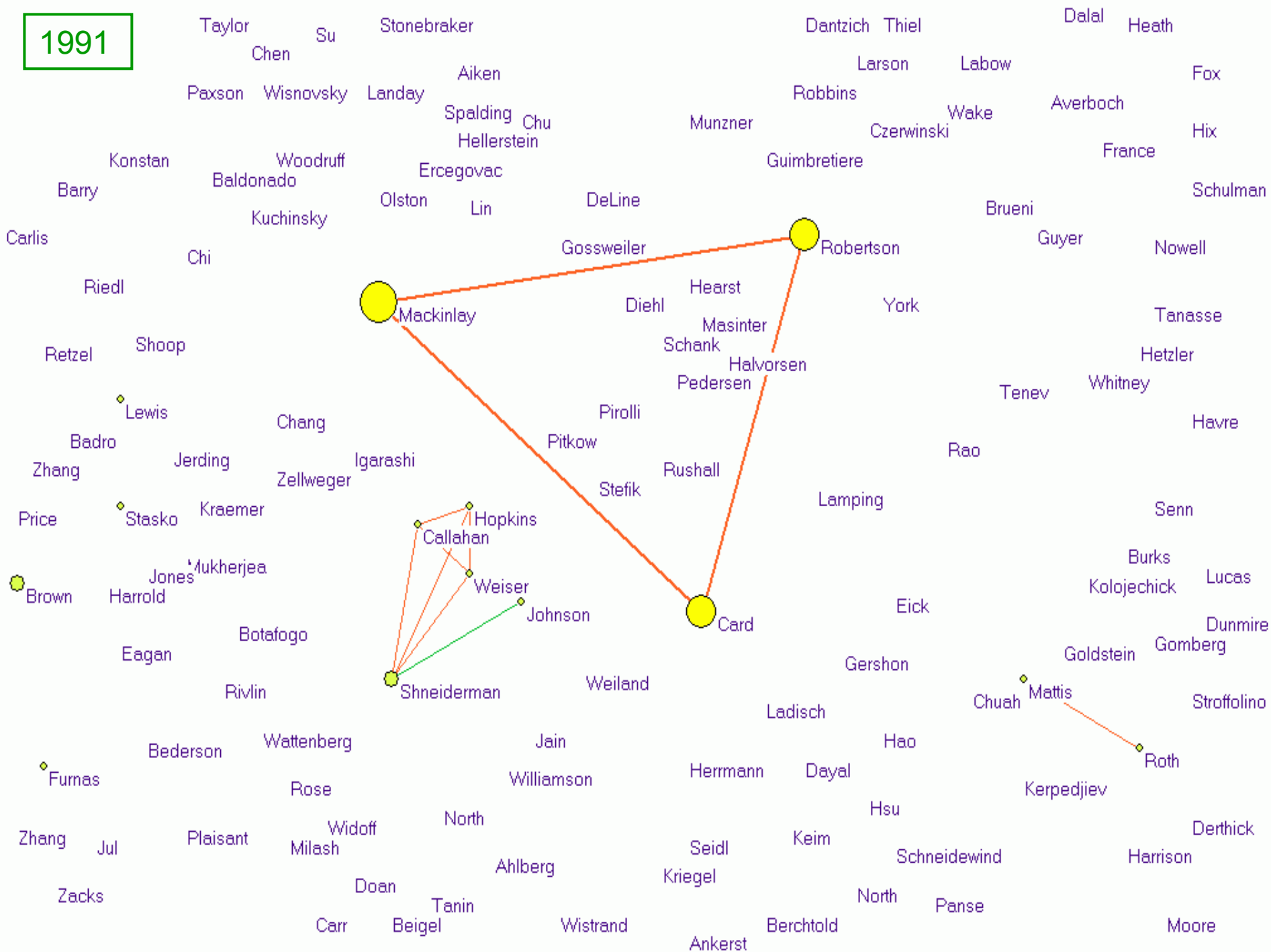




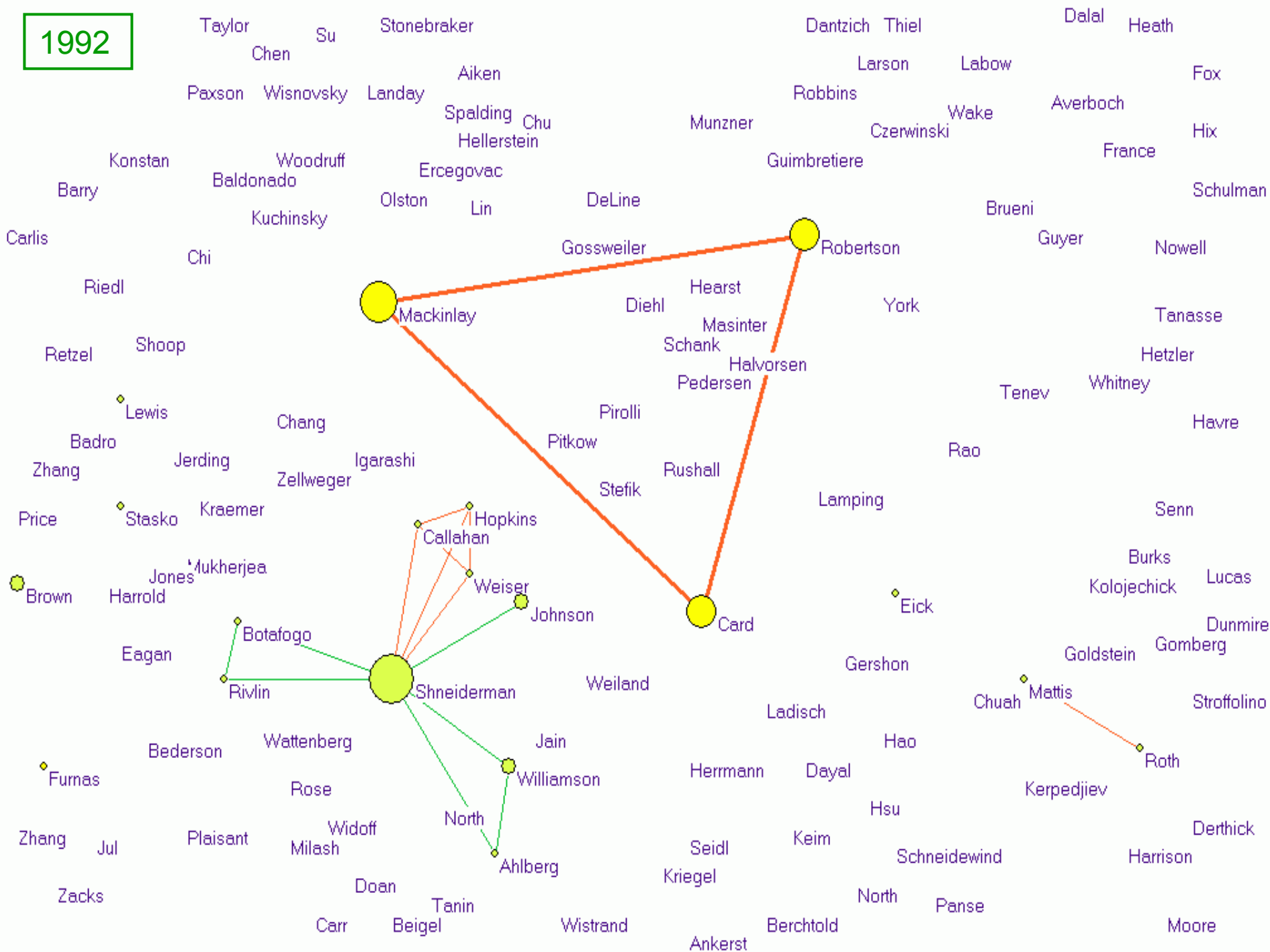
1990



1991

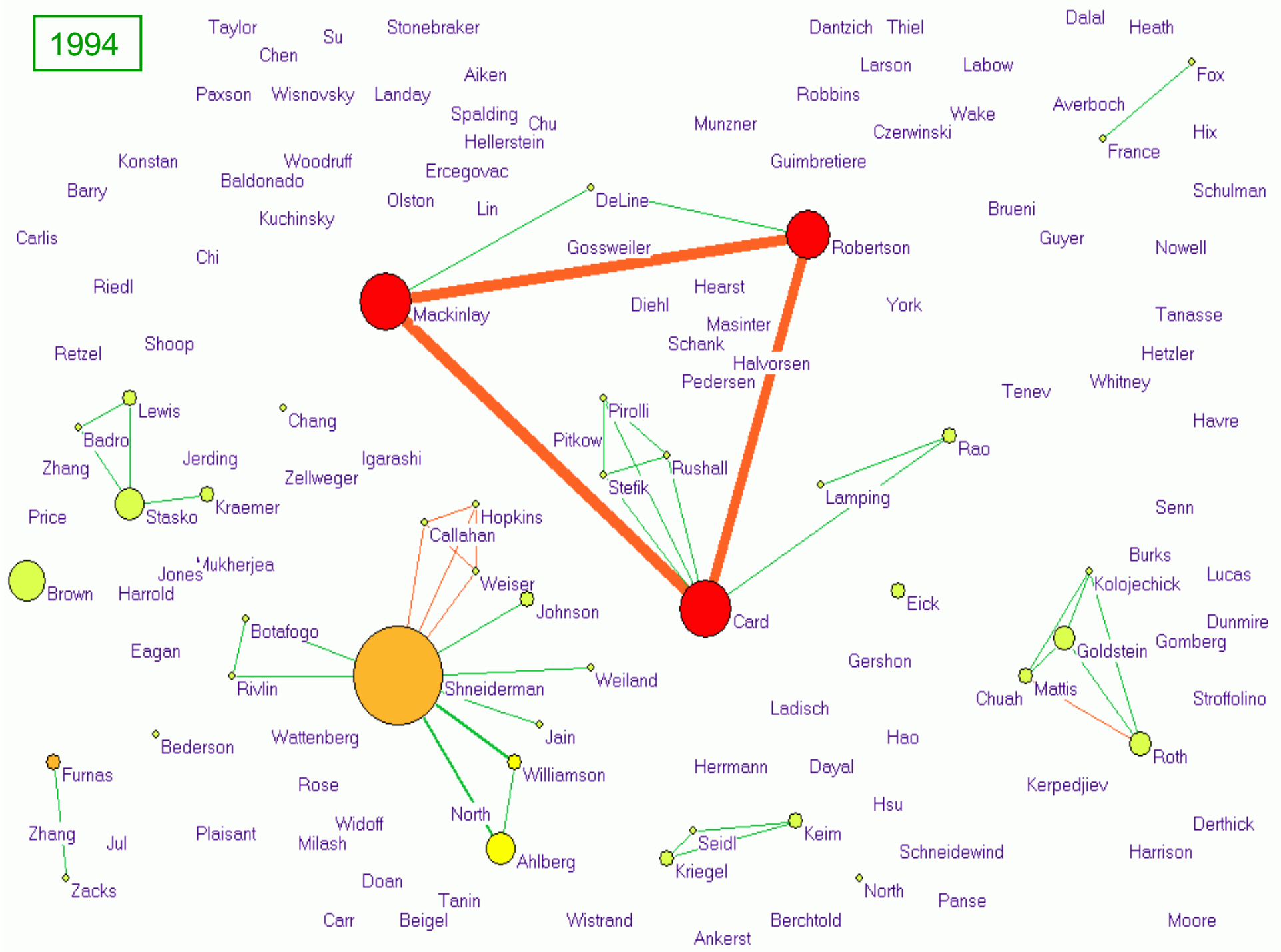


1992



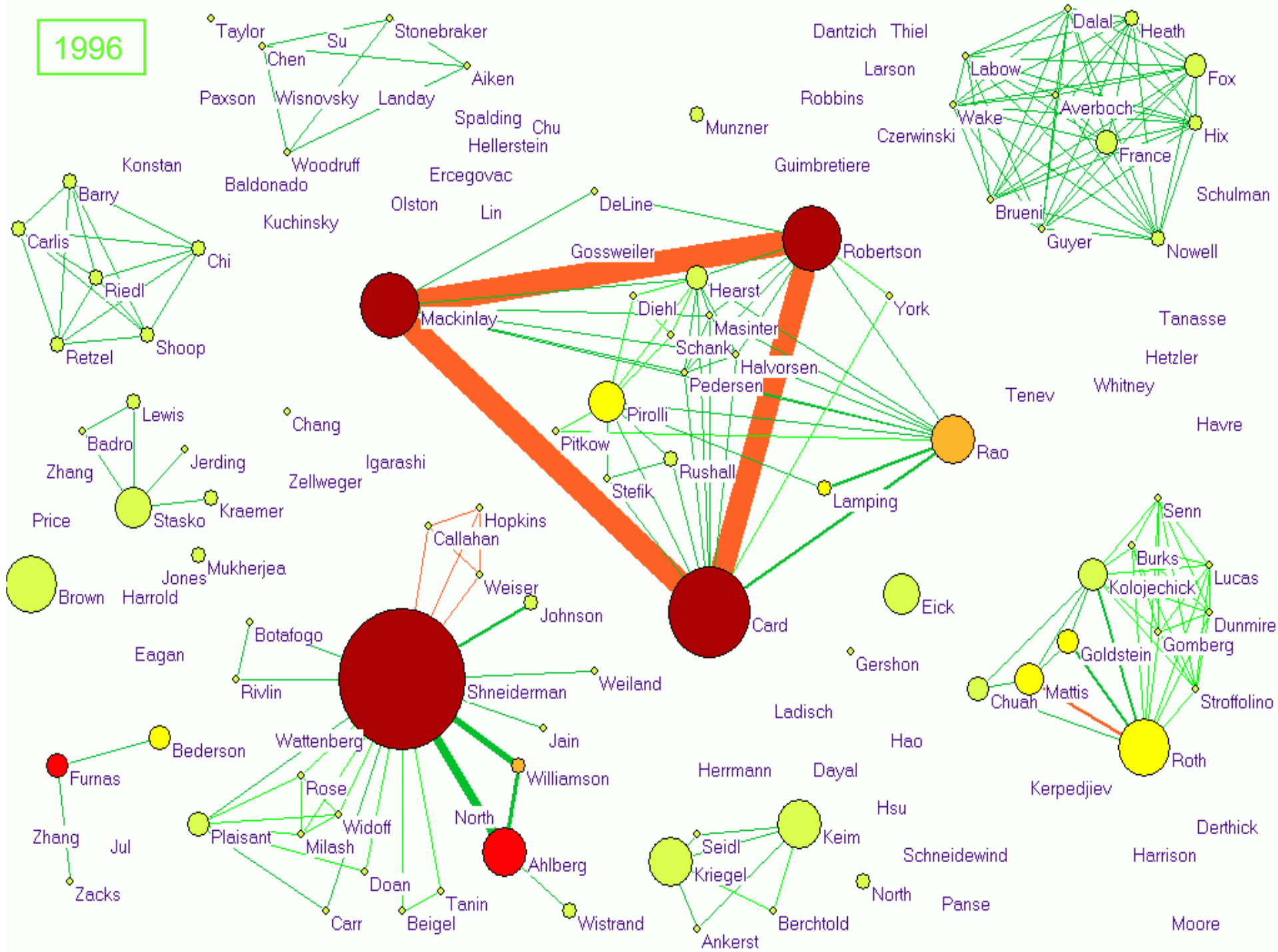


1994

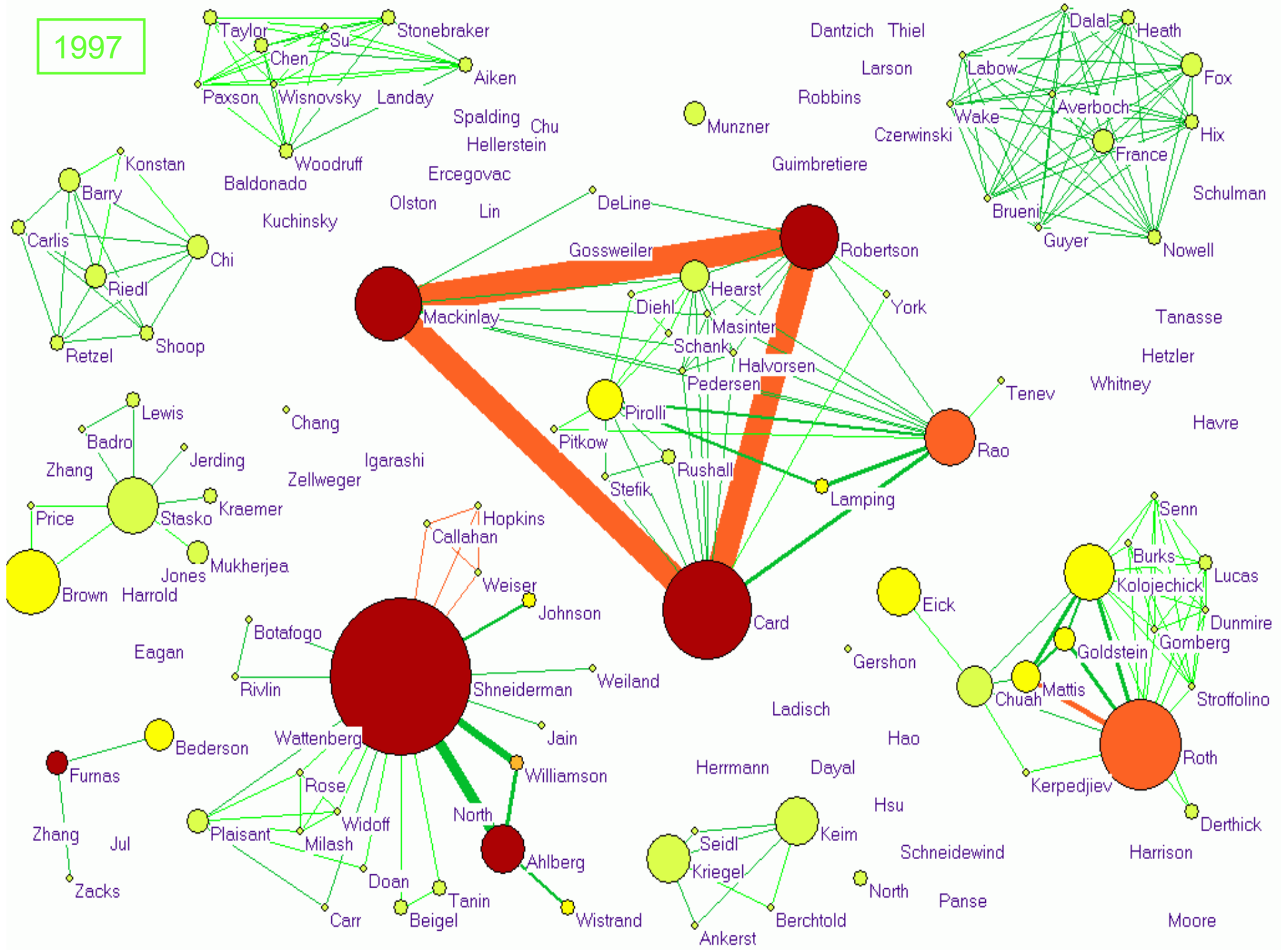




1996

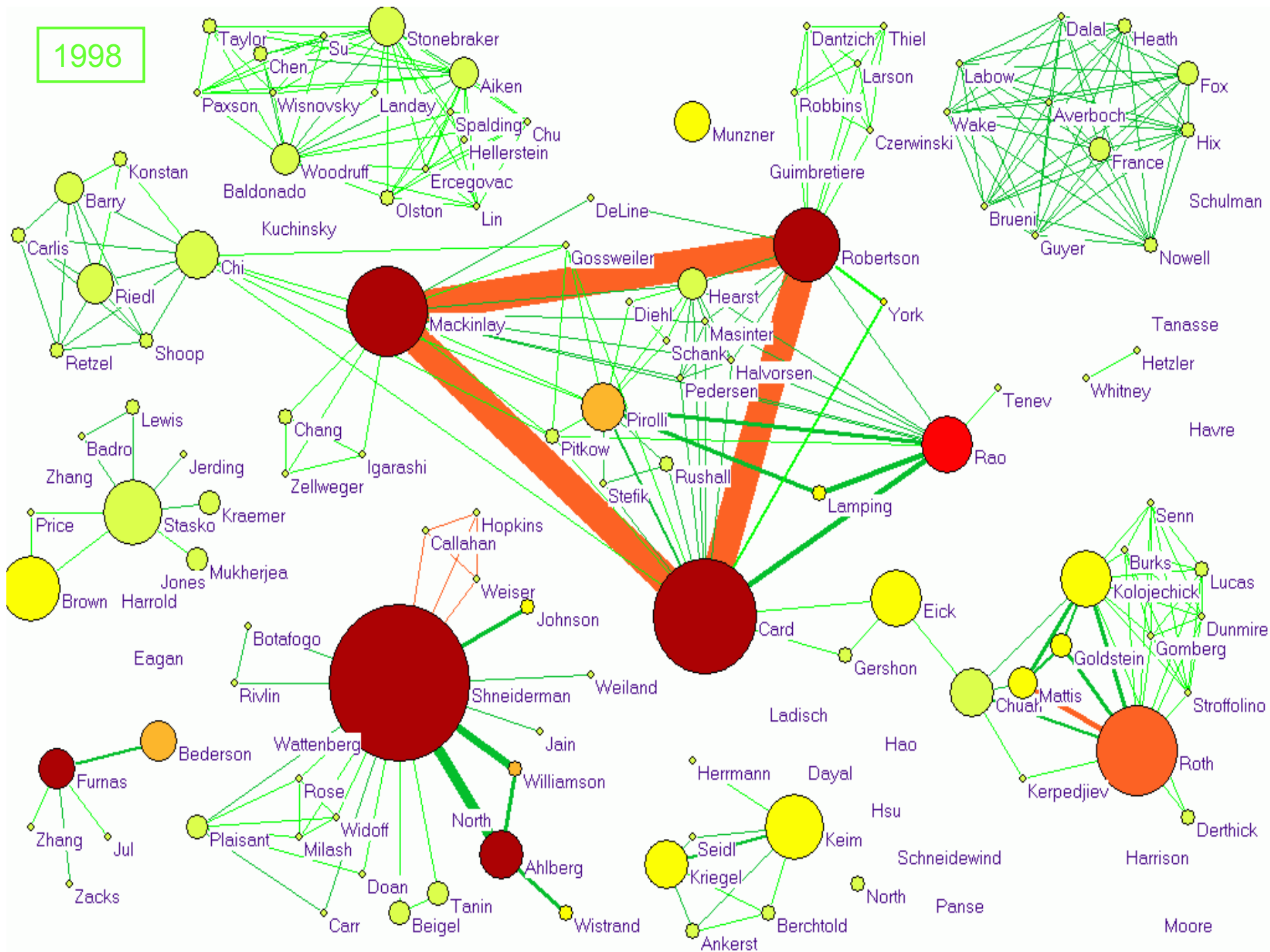


1997

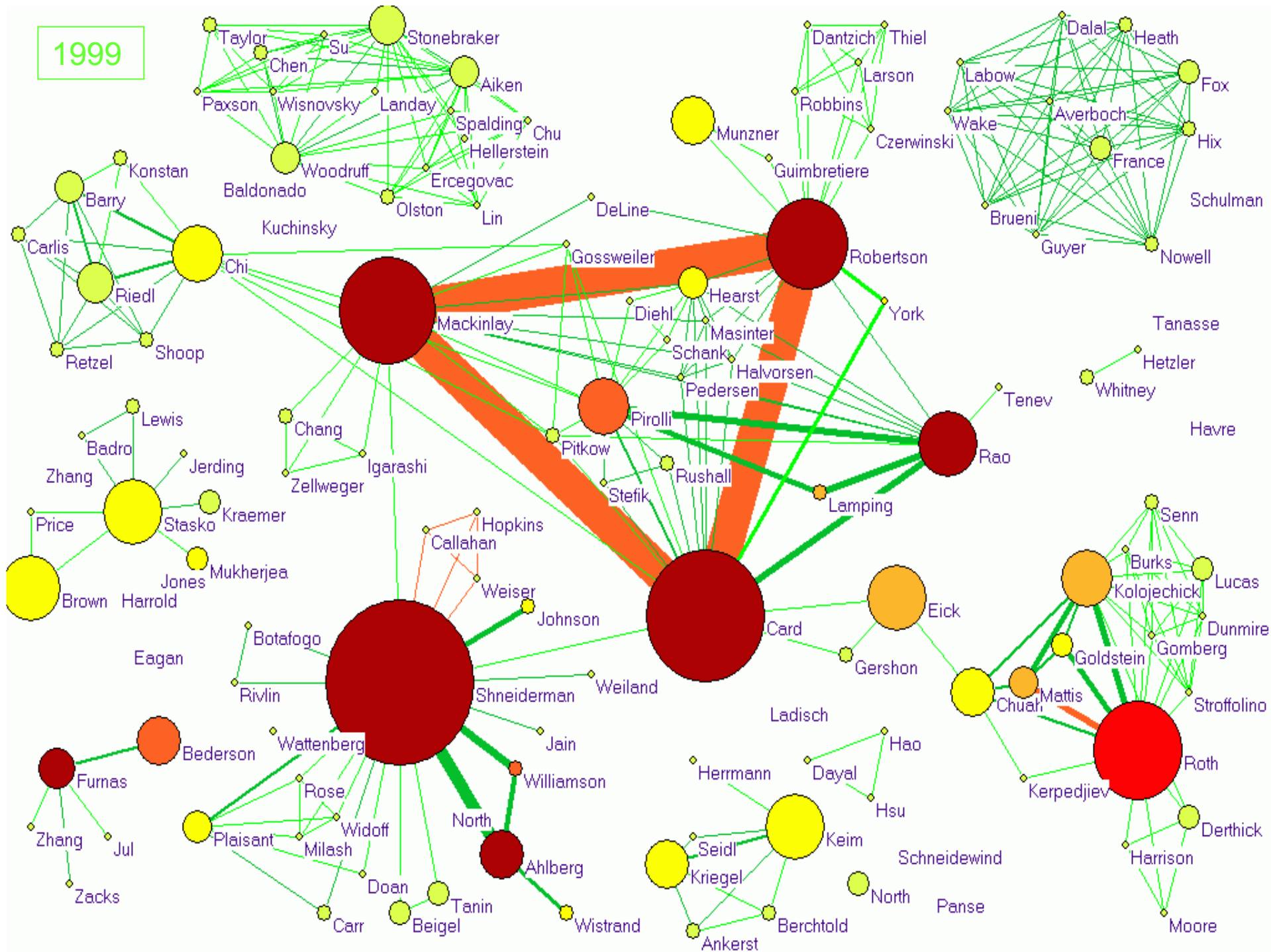




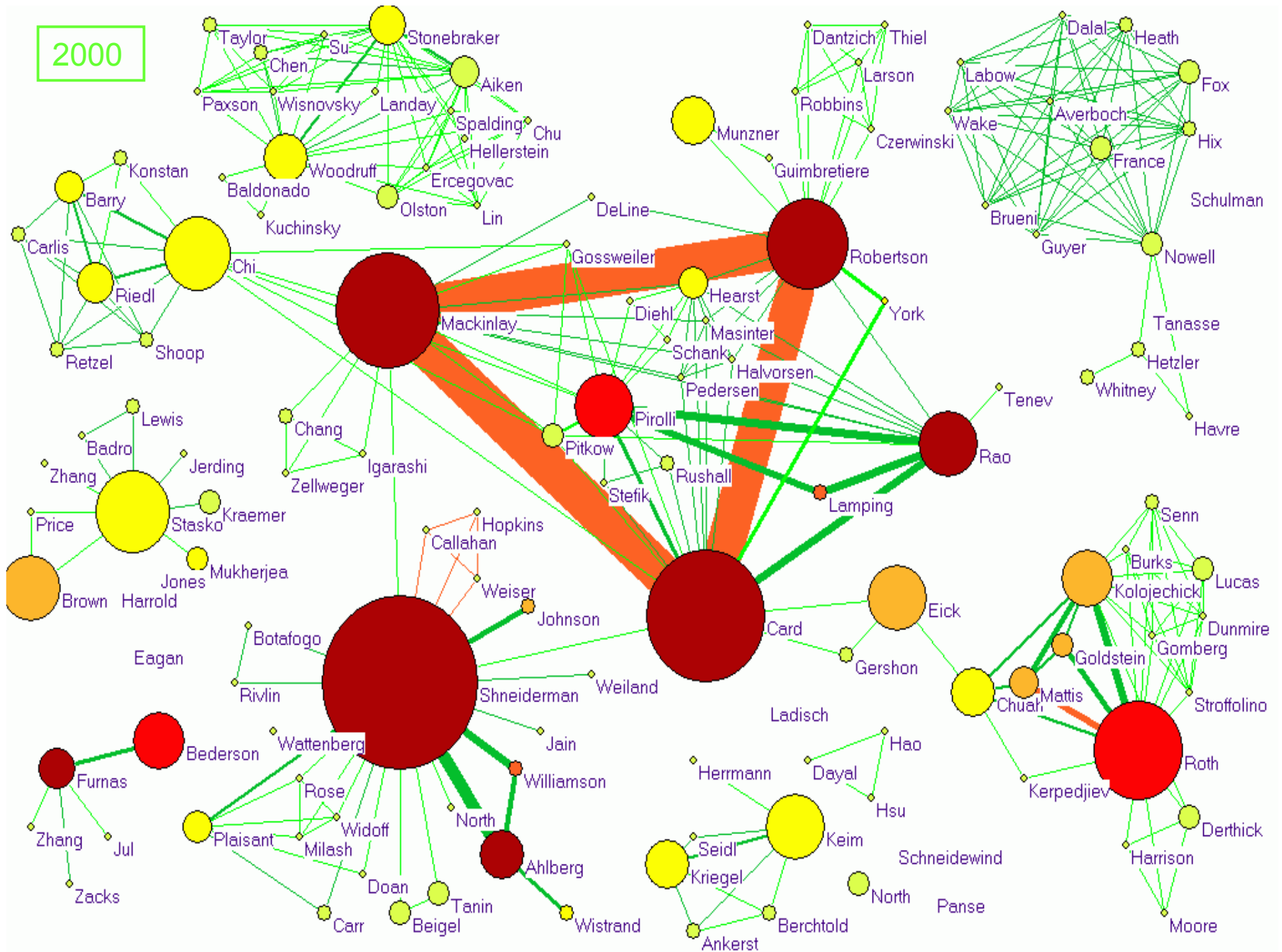
1998



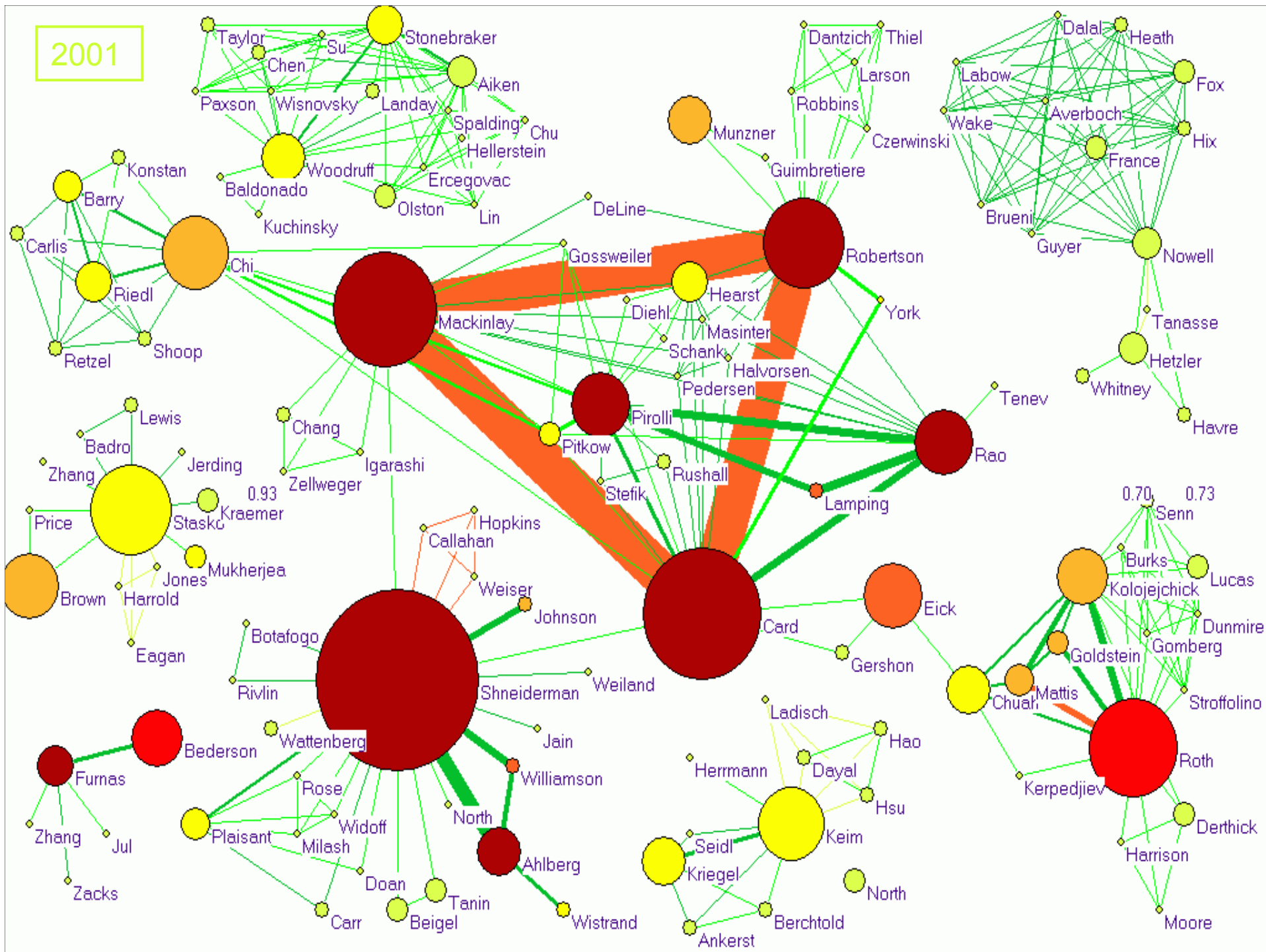
1999



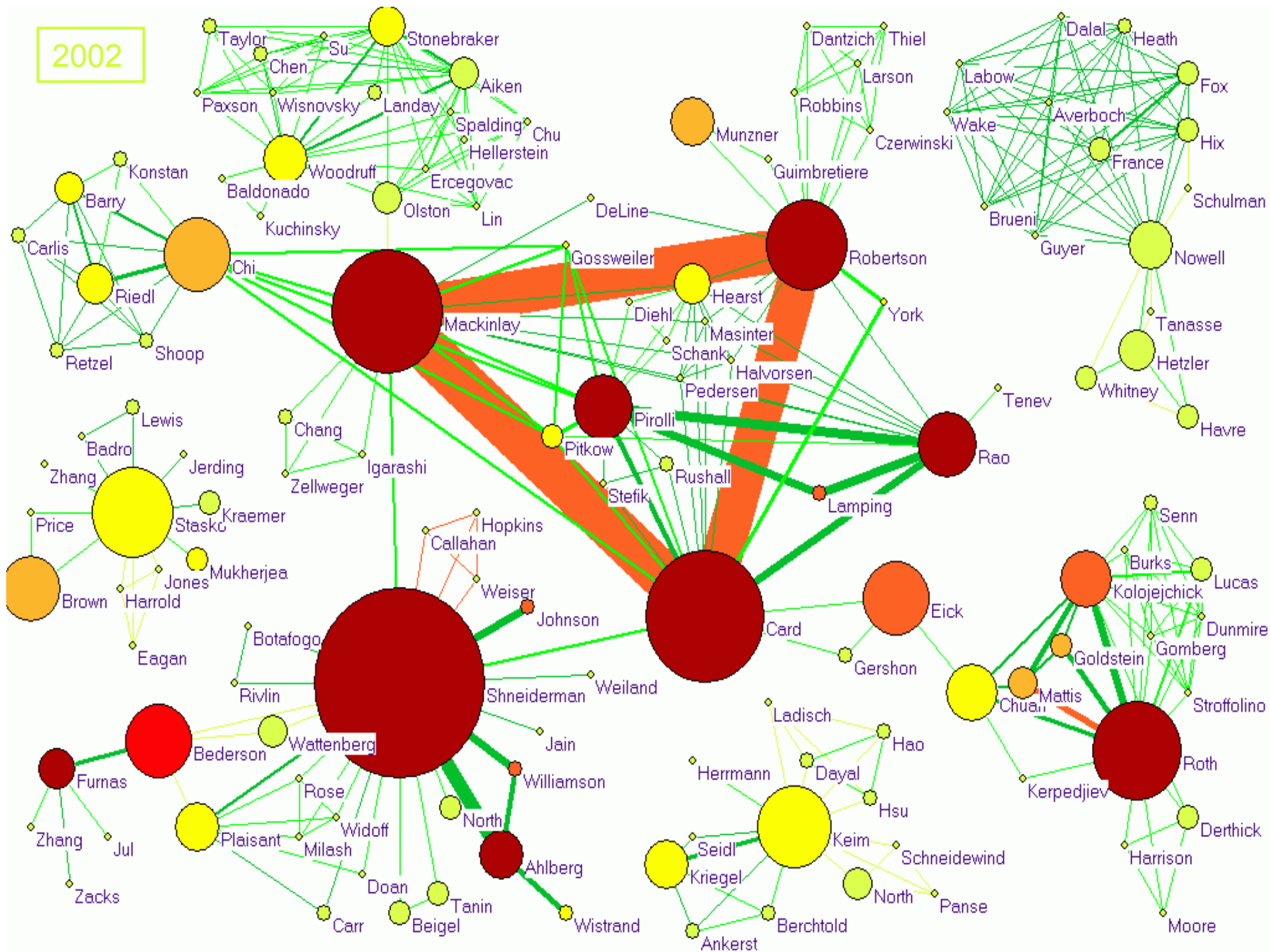
2000



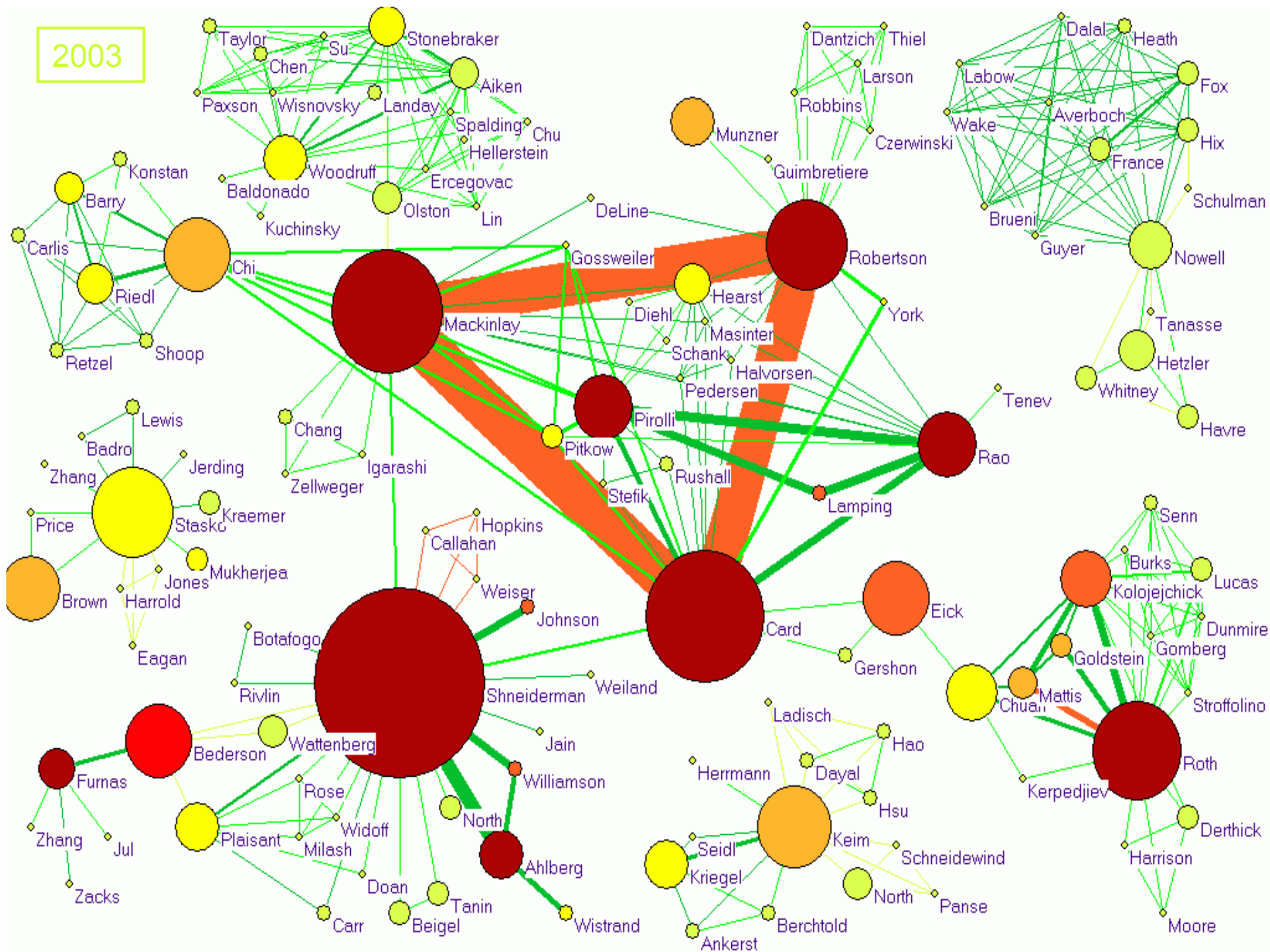
2001



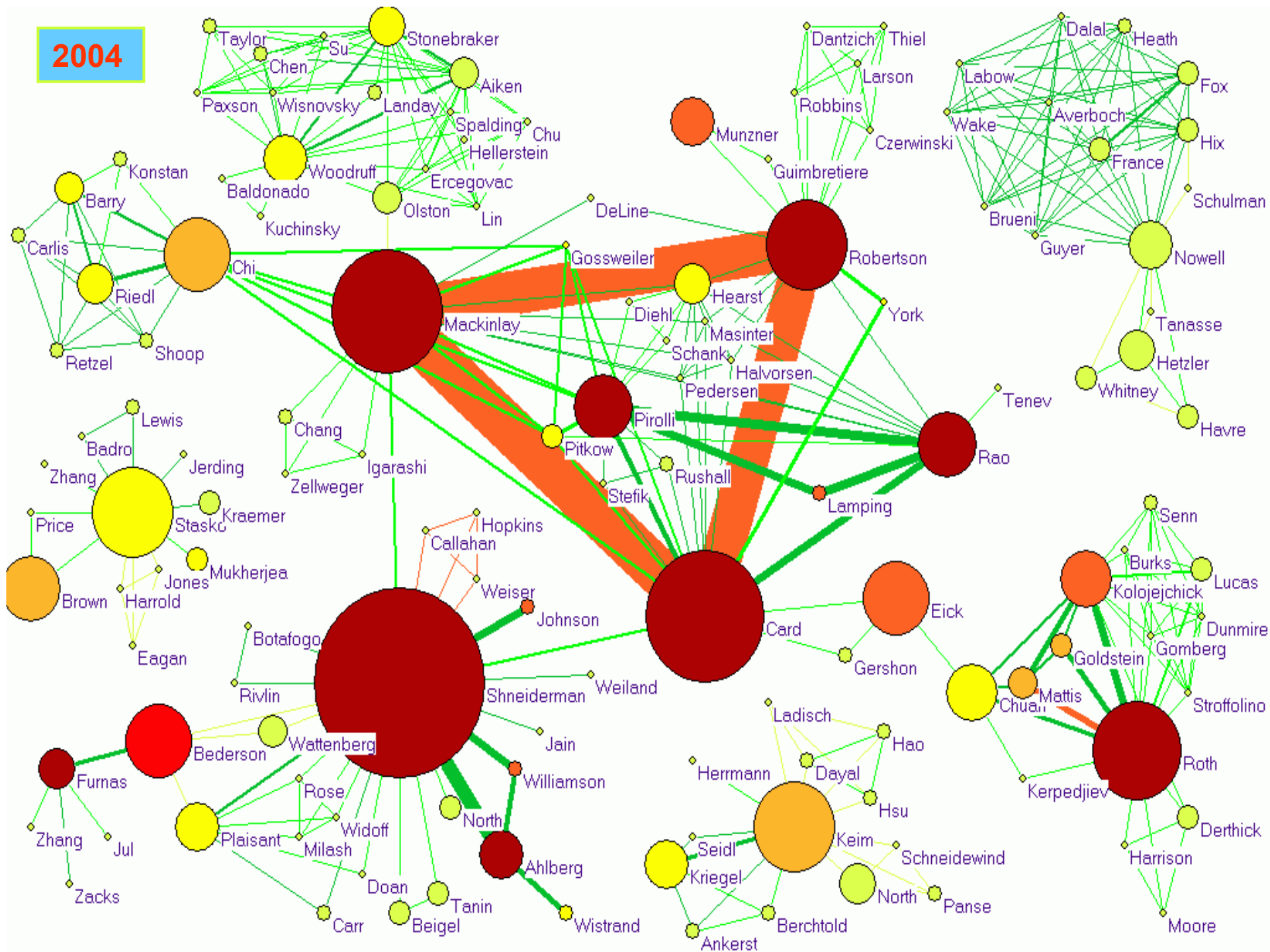
2002

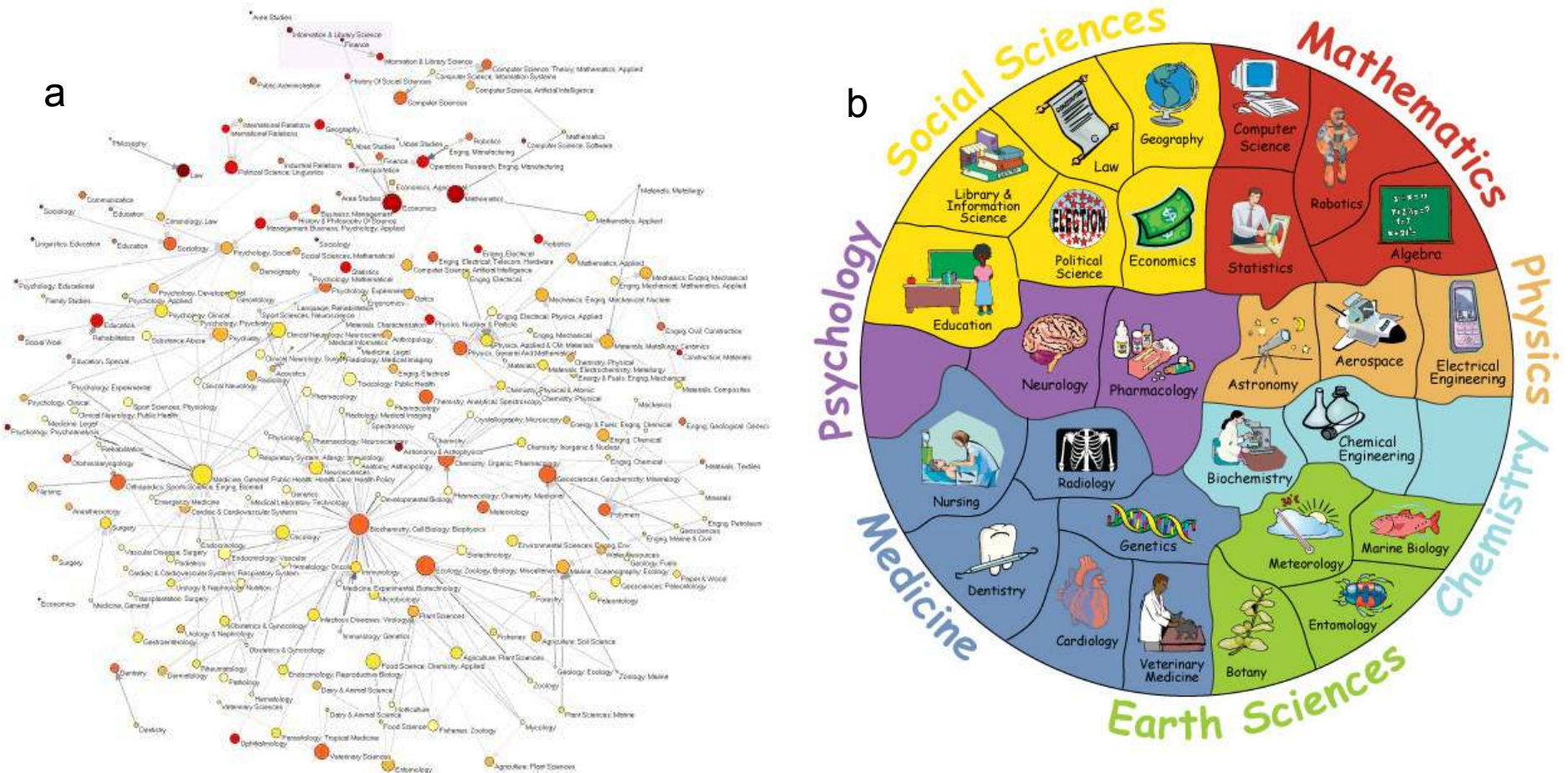


2003



2004





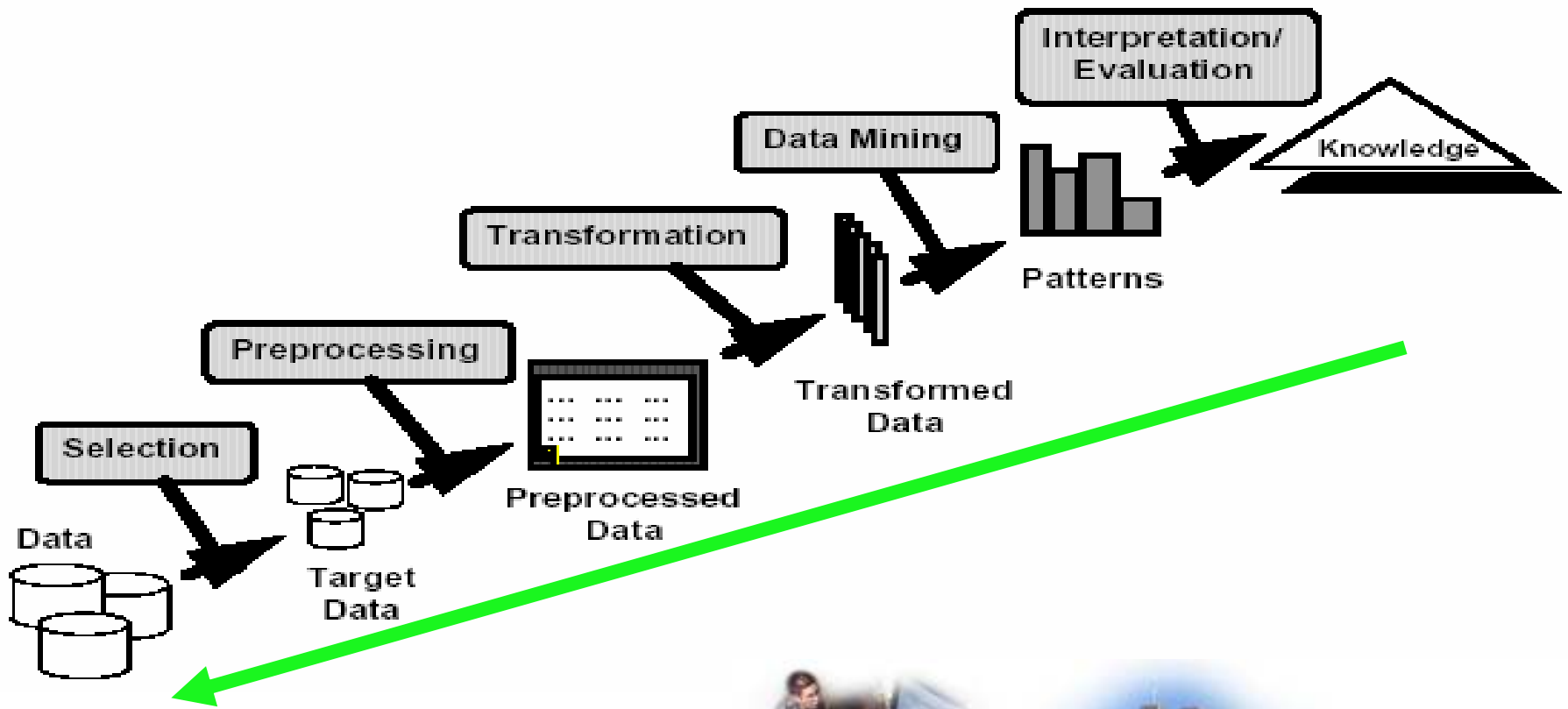
Map of Science derived from text mining of volumes of scientific literature. (a) Visualization for researchers where 2D proximity represents similarity of research areas and size and color of research areas indicate activity level. Interactive map allows drilling to research articles. (b) Puzzle for elementary school students that represents similarity of research areas.



**0308264**

**Vipin Kumar and Jaideep Srivastava, University of Minnesota  
Data Mining for Rare Class Analysis**

- A precursor to many attacks on networks is often a reconnaissance operation, more commonly referred to as a scan. These computers, once compromised, are used to send spam, serve pornography, or launch large scale attacks on the Internet.
- This project developed a new, effective method for scan detection by employing a data mining approach, which makes it possible to use automated techniques for knowledge discovery from massive data.
- Extensive experiments on real network traffic data have shown that the new methods have substantially better performance than the state of the art methods, in terms of coverage, false alarm rate and speed of detection.
- With millions of compromised computers constantly scanning the Internet, such improved scan detection techniques are critical to providing security analysts with the information they need to take preventive measures.



**IIS-0414380**

**Matthew O. Ward and Elke A. Rundensteiner, WPI**

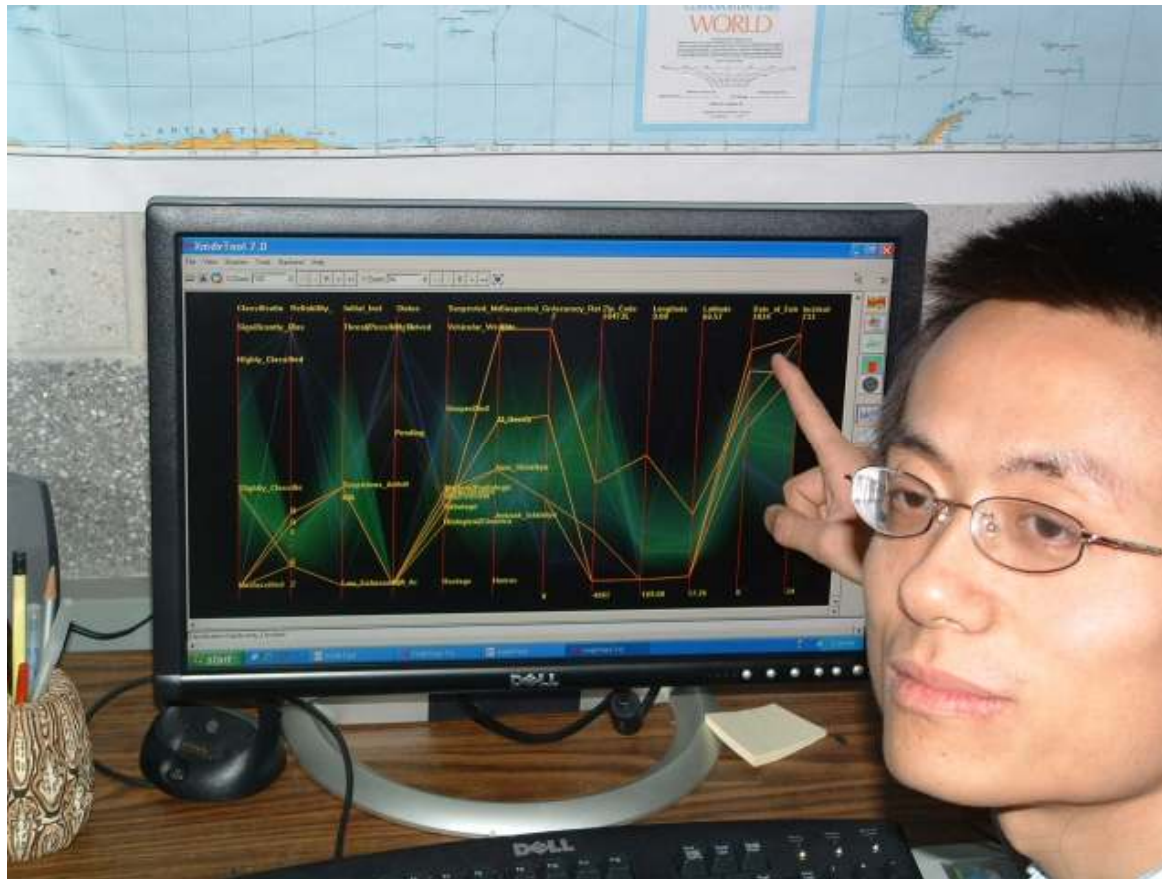
**Quality-Aware Visual Exploration Tools**

In general, exploratory data analysis assumes data is of high quality and that data transformations, either computational (e.g., data mining) or visualization, do not cause loss of information and result in new knowledge of high quality.

In reality, data is rarely of uniformly high quality or certainty, and every transformation performed distorts or loses information.

This disparity leads to results (new knowledge) that may be misleading or incorrect, with potentially serious consequences.

This project adds a new and important dimension to knowledge discovery – making the researcher aware of the quality of the data they are analyzing as well as the information loss resulting from filtering, sampling, clustering, visualizing, and all other transformations applied to the data, resulting in more accurate and reliable knowledge and decisions can be drawn from data.



A data analyst points out high confidence clusters in a simulated homeland security dataset. Lower quality information has been automatically deemphasized.

# Opportunities at NSF: <http://www.nsf.gov>

- **MyNSF**
  - Create your profile
  - Receive
    - Announcement on funding opportunities
      - **Explosives and Related Threats: Frontiers in Prediction and Detection (EXP)**
    - Relevant News, Discoveries, reports, ...
- **Find Funding**
  - Search by keywords, dates, directorates, ...
  - Browse the NSF Web:
    - **NSF-wide/Cross-cutting Programs**
      - Integrative Graduate Education and Research Traineeship Program (IGERT)
      - NSF Graduate Teaching Fellowship in K-12 Education (GK-12)
      - Grant Opportunities for Academic Liaison with Industry (GOALI)
    - **Office of International Science & Engineering**
      - Partnerships for International Research and Education (PIRE)
- **Awards Search**
  - Search by keywords
    - Find researchers, relevant NSF programs, program directors, ...
  - Search by NSF Units (divisions, programs, ...)
    - Get an idea about the scope

# CISE++ Opportunities:

<http://www.nsf.gov/dir/index.jsp?org=CISE>

Cyber-Enabled Discovery and Innovation (CDI)

Information & Intelligent Systems Program Solicitation

Expeditions in Computing

Foundations of Data and Visual Analytics

CreativeIT

CyberTrust

Community-Based Data Interoperability Networks

Sustainable Digital Data Preservation and Access Network Partners (DataNet)

Mathematical Sciences: Innovations at the Interface with Computer Sciences

Industry/University Cooperative Research Centers Program (I/UCRC)

CISE Pathways to Revitalized Undergraduate Computing Education (CPATH )

Research Experience for Undergraduates (REU) Sites & Supplements