Dynamic and Social Network Analysis

Lecture 12

Miray Kas
Bilkent University
Computer Engineering Department

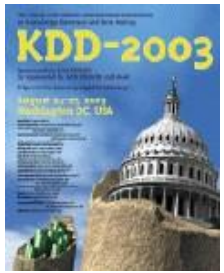
Authorship, Science, Citation Networks

KDD Cup 2003 Dataset

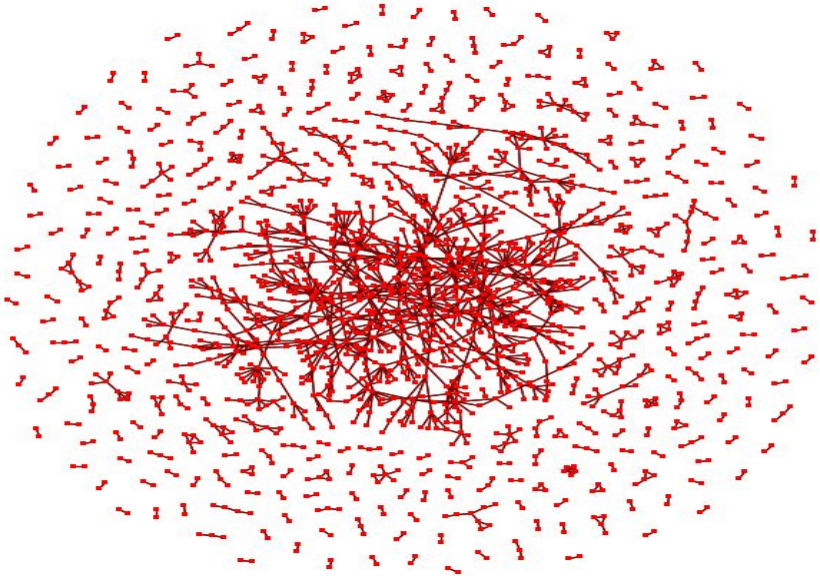
- The data consists of roughly 29,000 papers in arXiv (1992-2003), within the field of high-energy physics.
- The data is structured as follows:
- ***(i)*** LaTeX sources of each paper (classified by year)
- ***(ii)*** abstract of each paper
- ***(iii)*** Date of each paper
- ***(iv)*** Citation graph data in the form of (citing_paper_id, cited_paper_id).
- https://www.cs.cornell.edu/projects/kddcup/datasets.html

KDD Cup 2003 Dataset

Authors	Papers	Citations	Publications	Co-authorship
8K	27K	352K	55K	38K

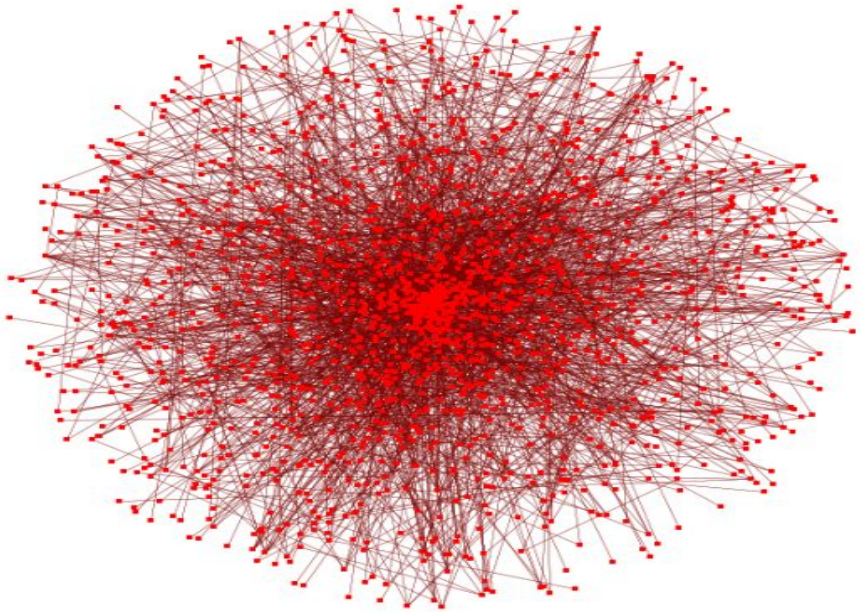


Co-publication/Author to Author Graph



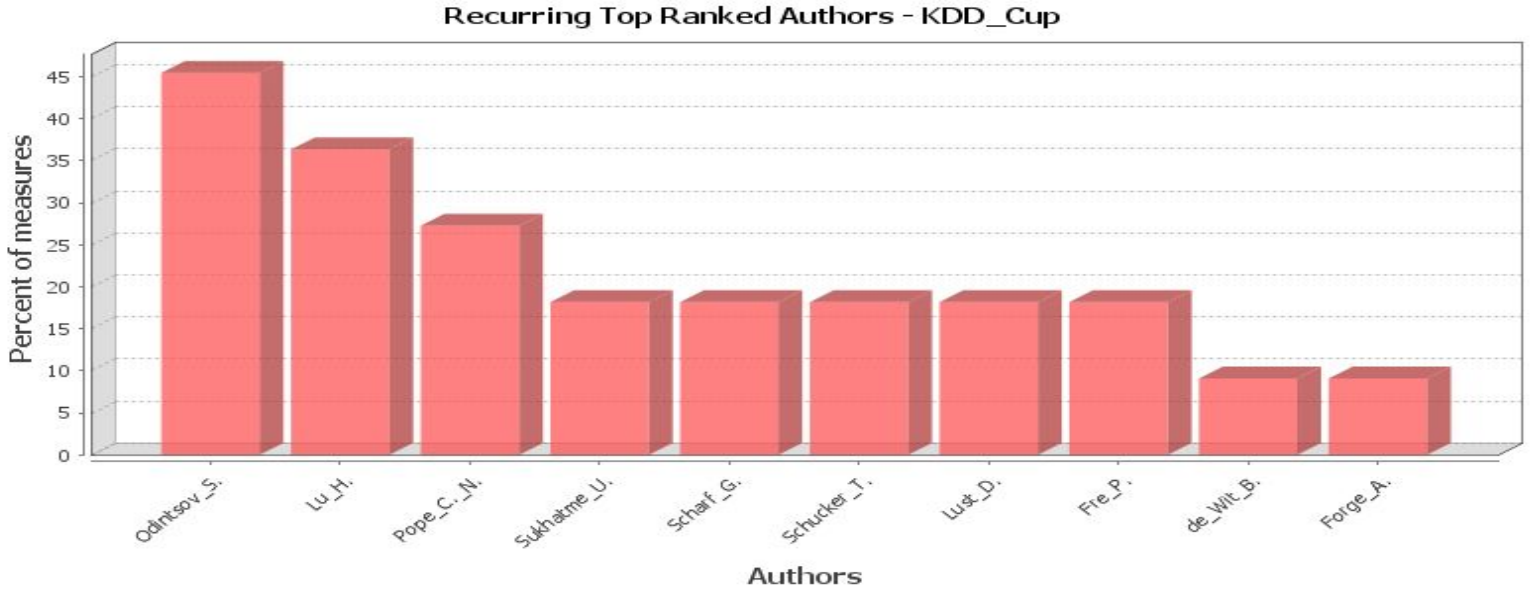
Co-Authorship Network

- Significant Dyads and Triads
- High Impact of Social Network



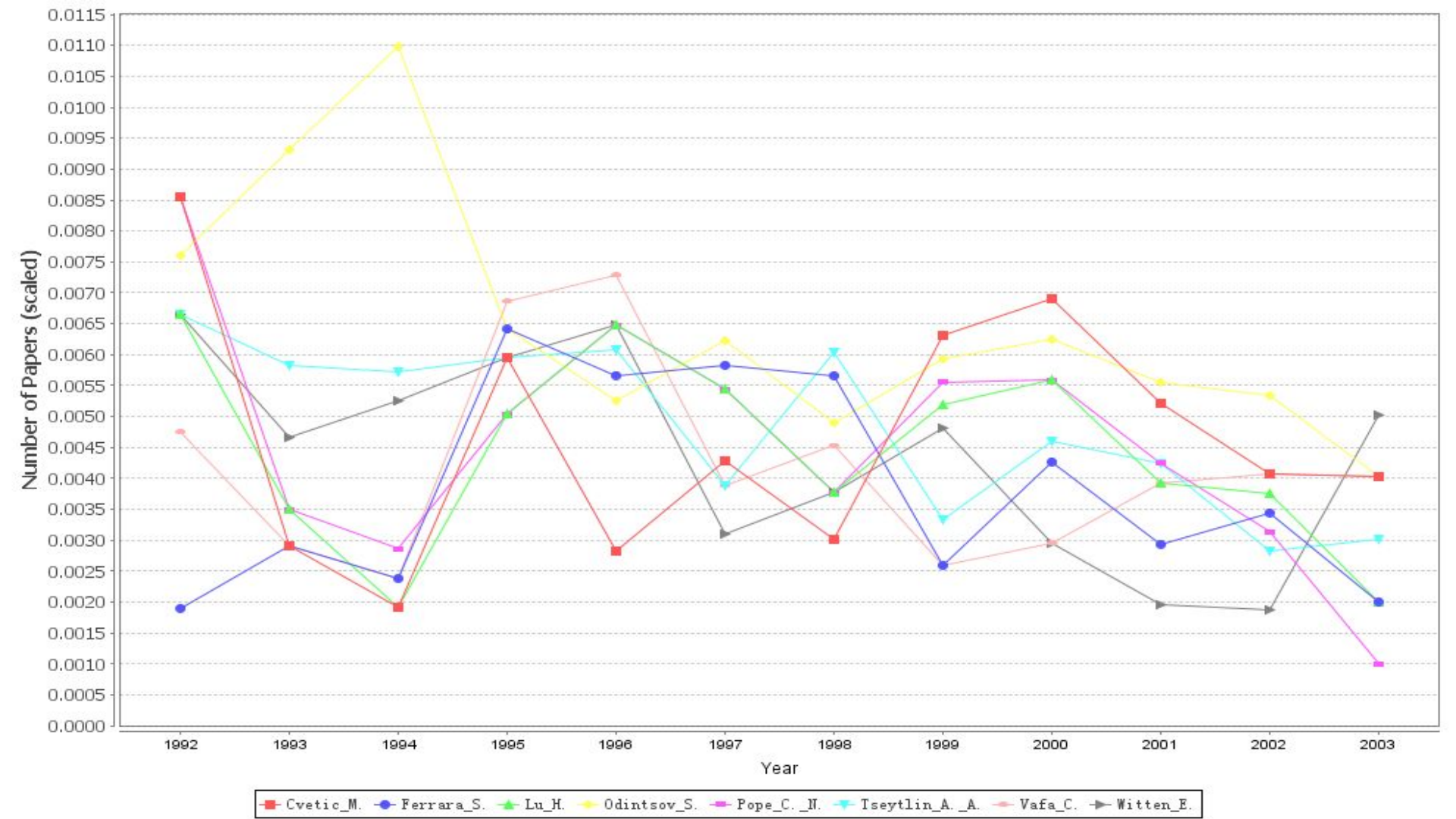
Co-Authorship & Citation Combined Network

Co-publication/Author to Author Graph

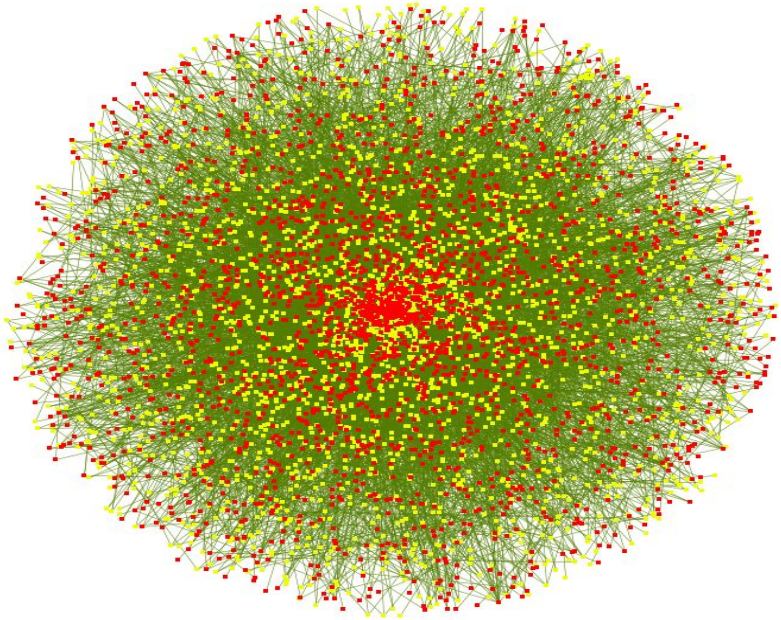


Author Attributes Analysis

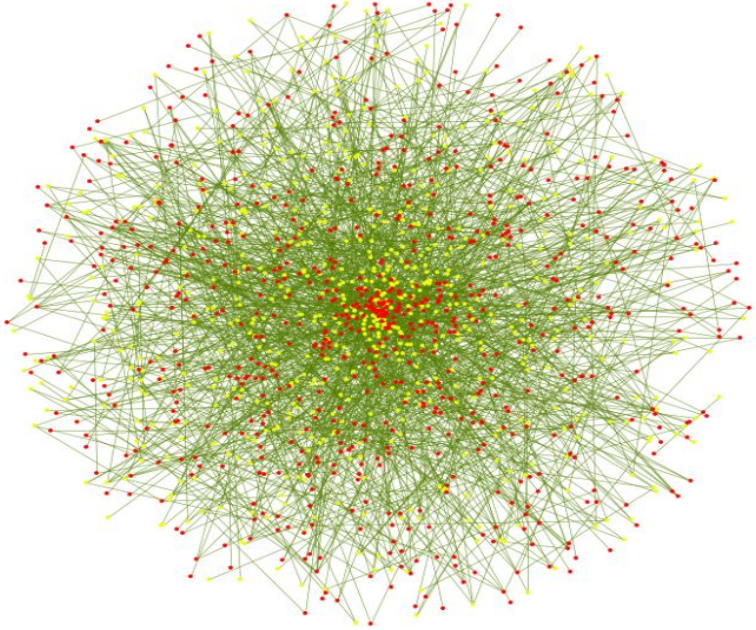
Rank	Author	Number of Papers
1	Odintsov_S.	176
2	Tseytlin_A._A.	131
3	Cvetic_M.	128
4	Pope_C._N.	128
5	Lu_H.	125
6	Vafa_C.	115
7	Ferrara_S.	112
8	Witten_E.	112


Number of Papers (scaled)
0.0115
0.0110
0.0105
0.0100
0.0095
0.0090
0.0085
0.0080
0.0075
0.0070
0.0065
0.0060
0.0055
0.0050
0.0045
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
Year
Cvetic_M.
Ferrara_S.
Lu_H.
Odintsov_S.
Pope_C._N.
Tseytlin_A._A.
Vafa_C.
Witten_E.

Author to Paper Analysis

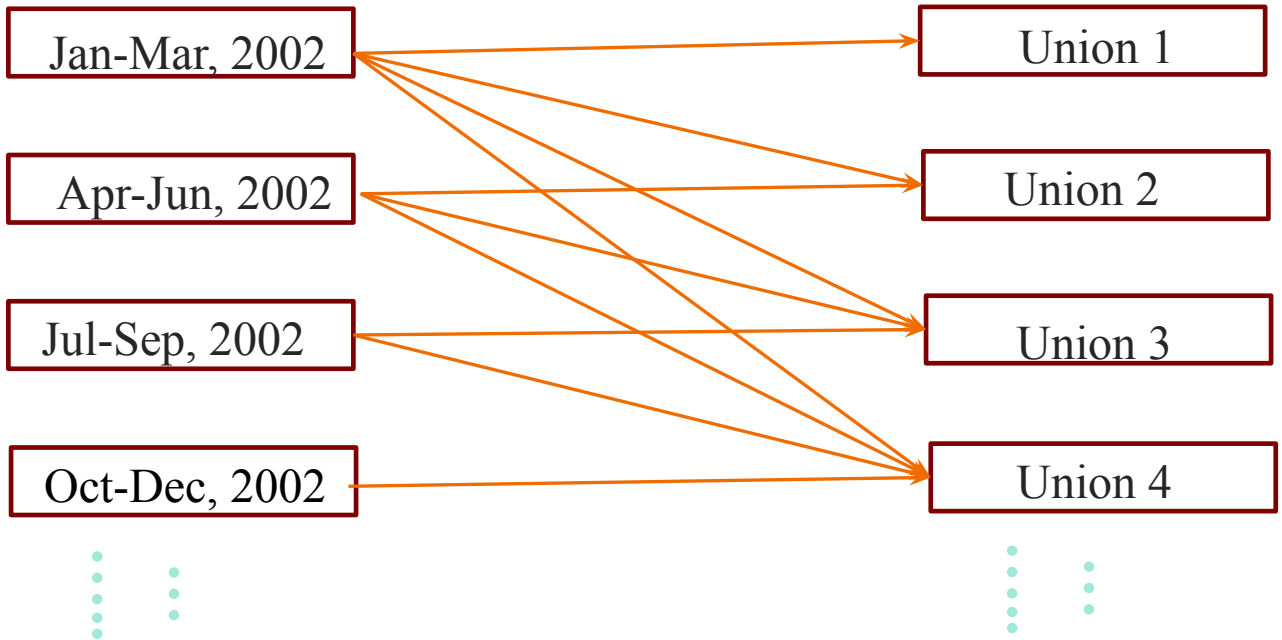


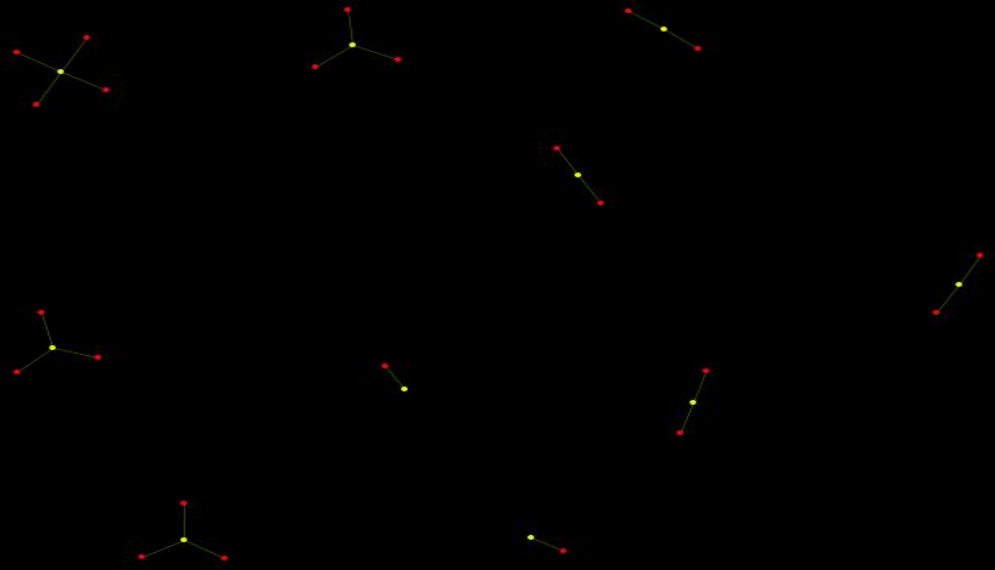
Before Data Cleaning

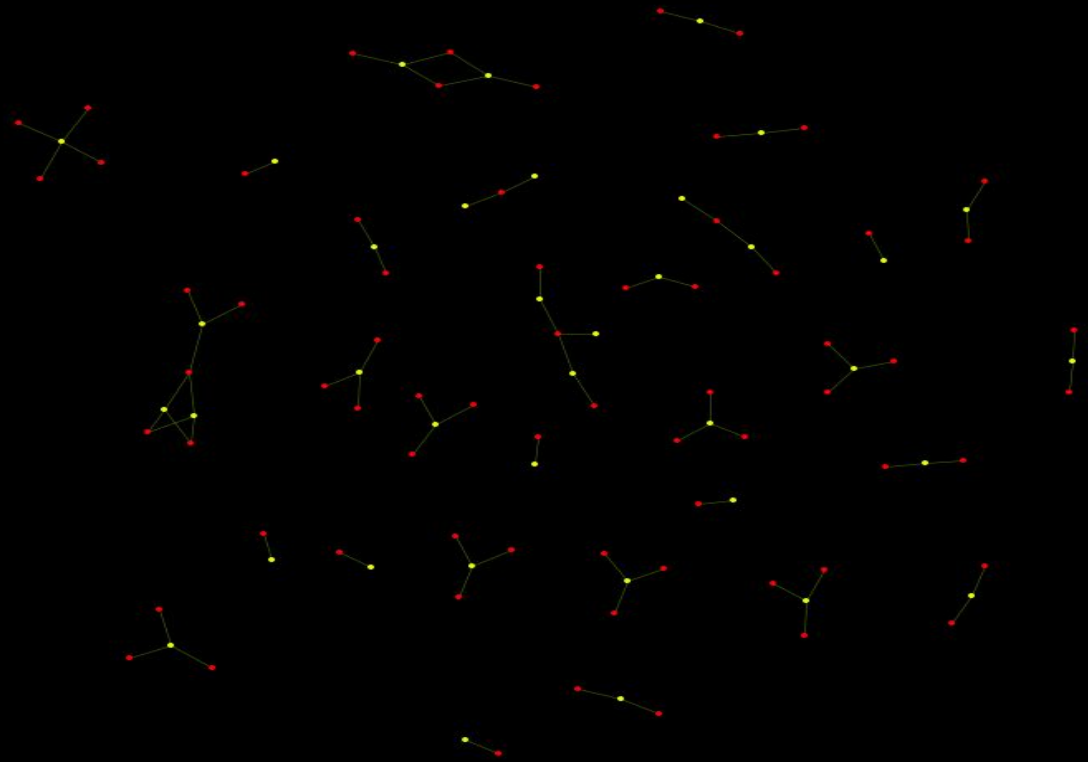


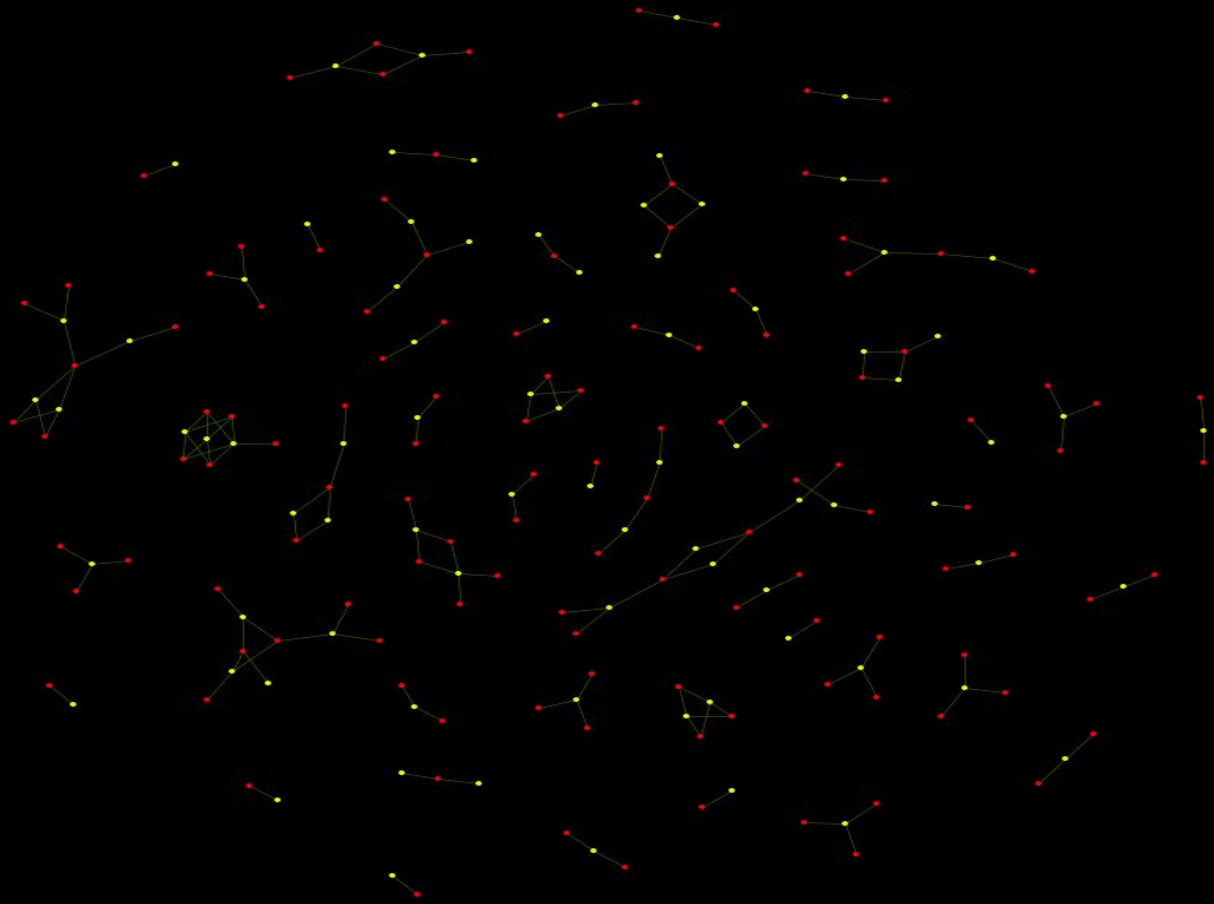
After Data Cleaning

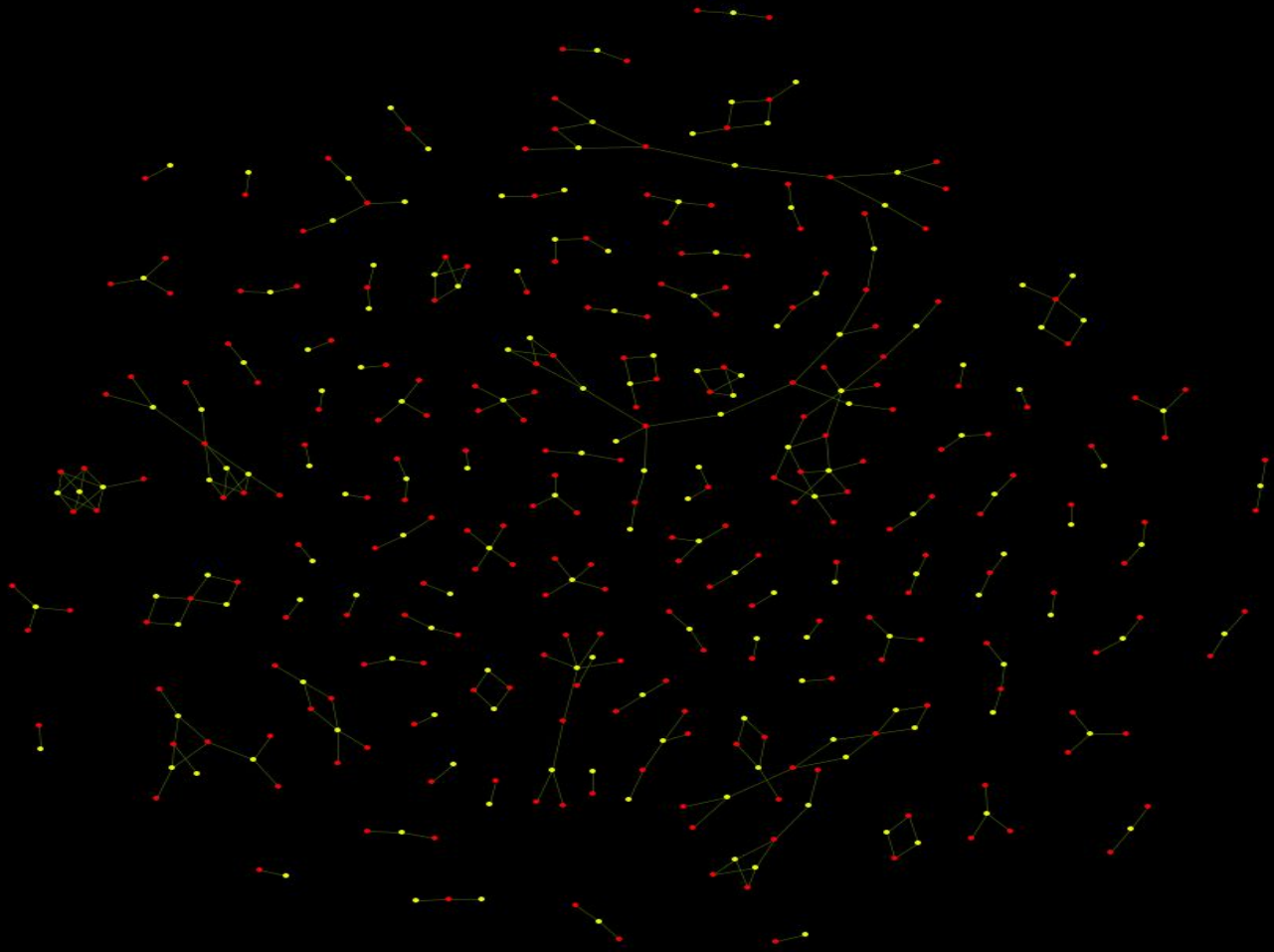
Transform Separate Date to Union Data

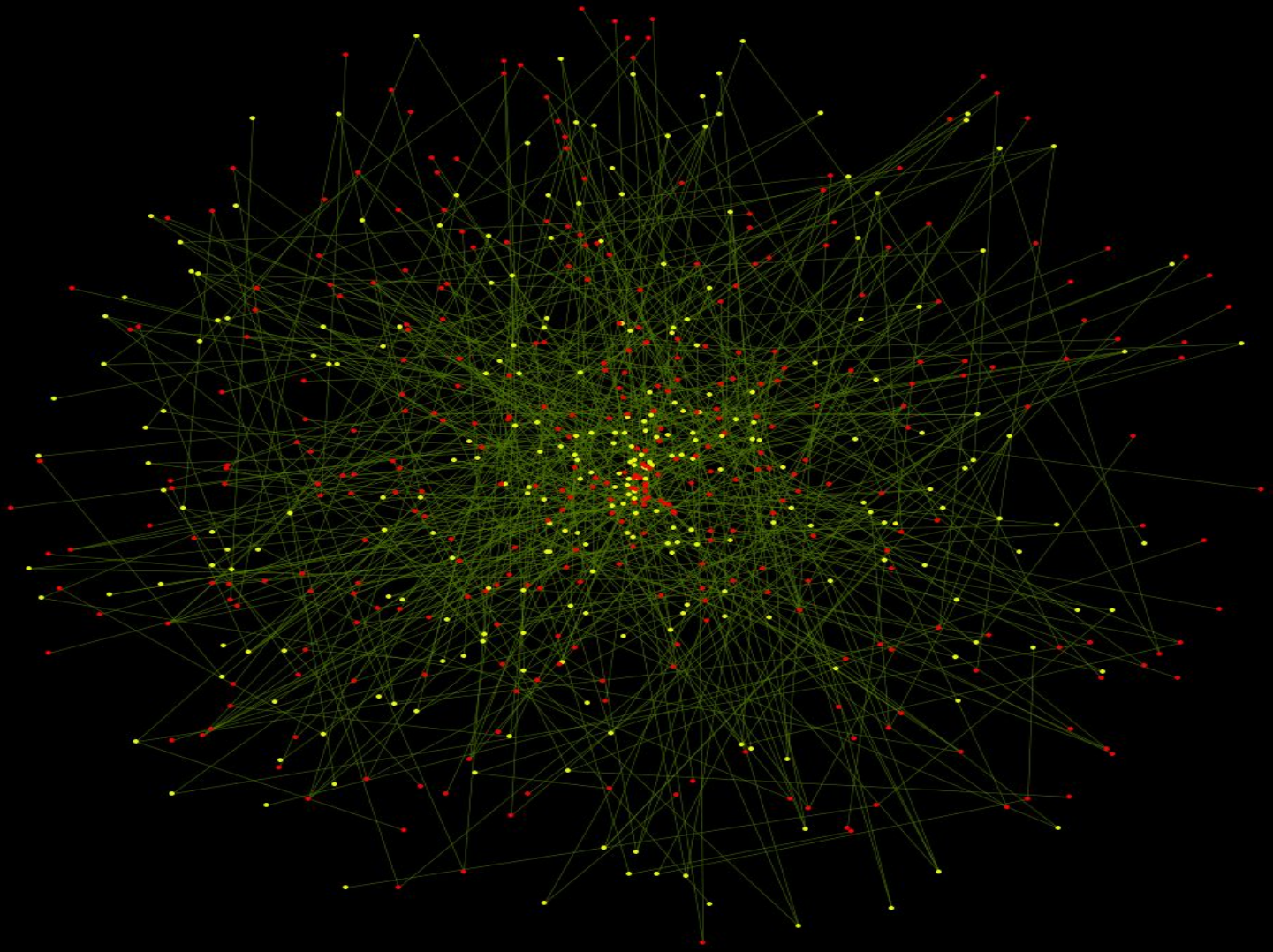


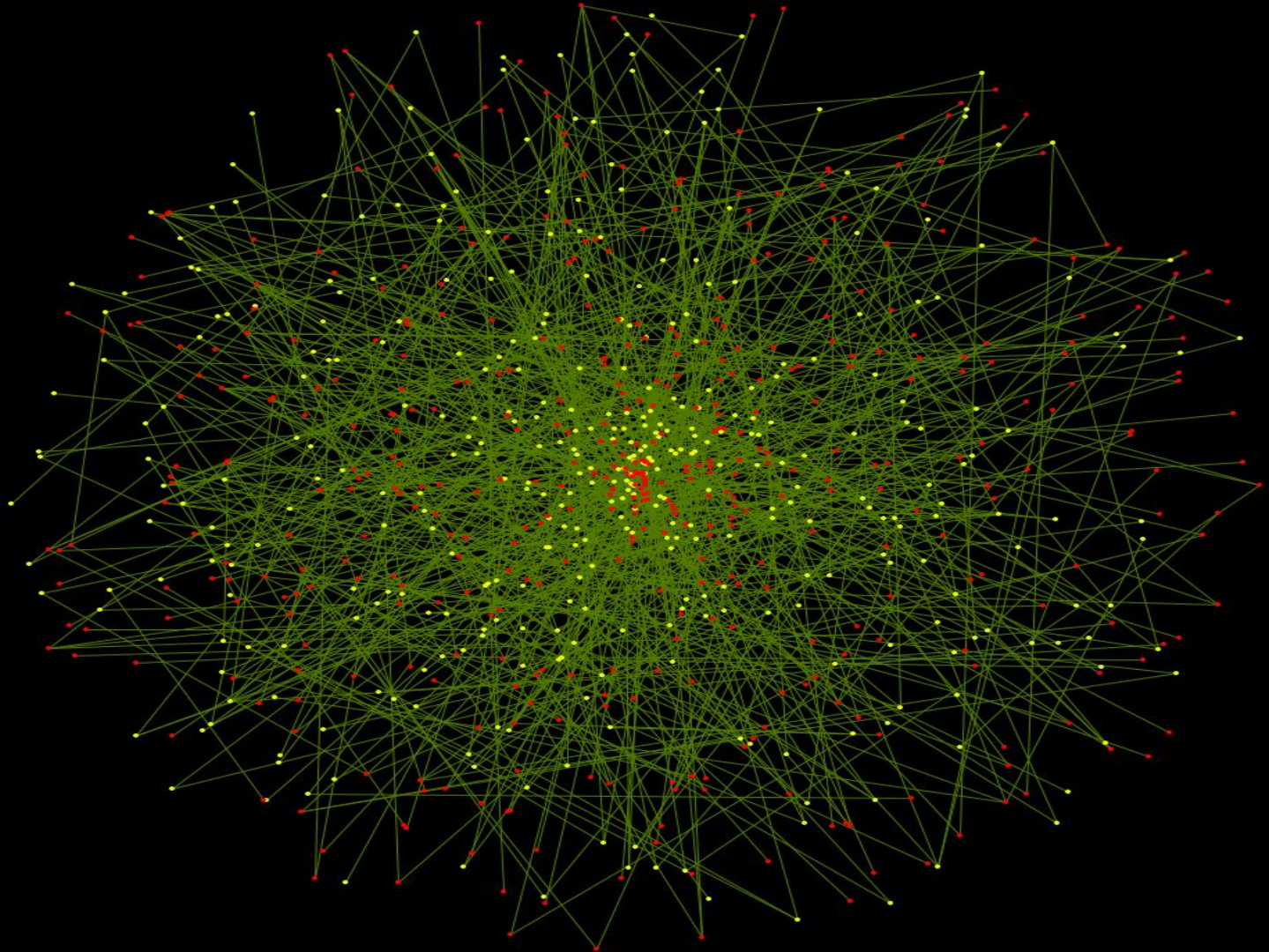


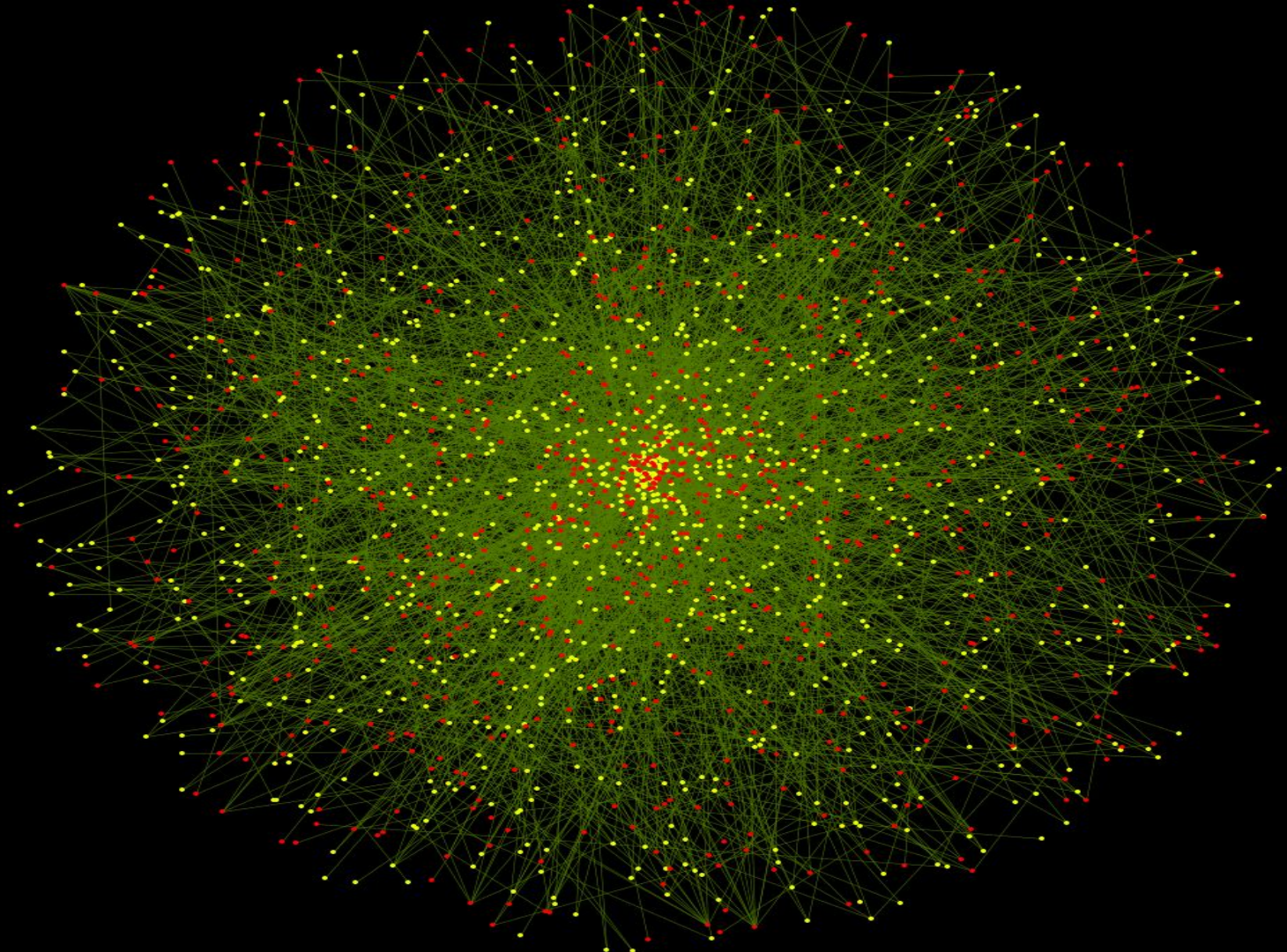


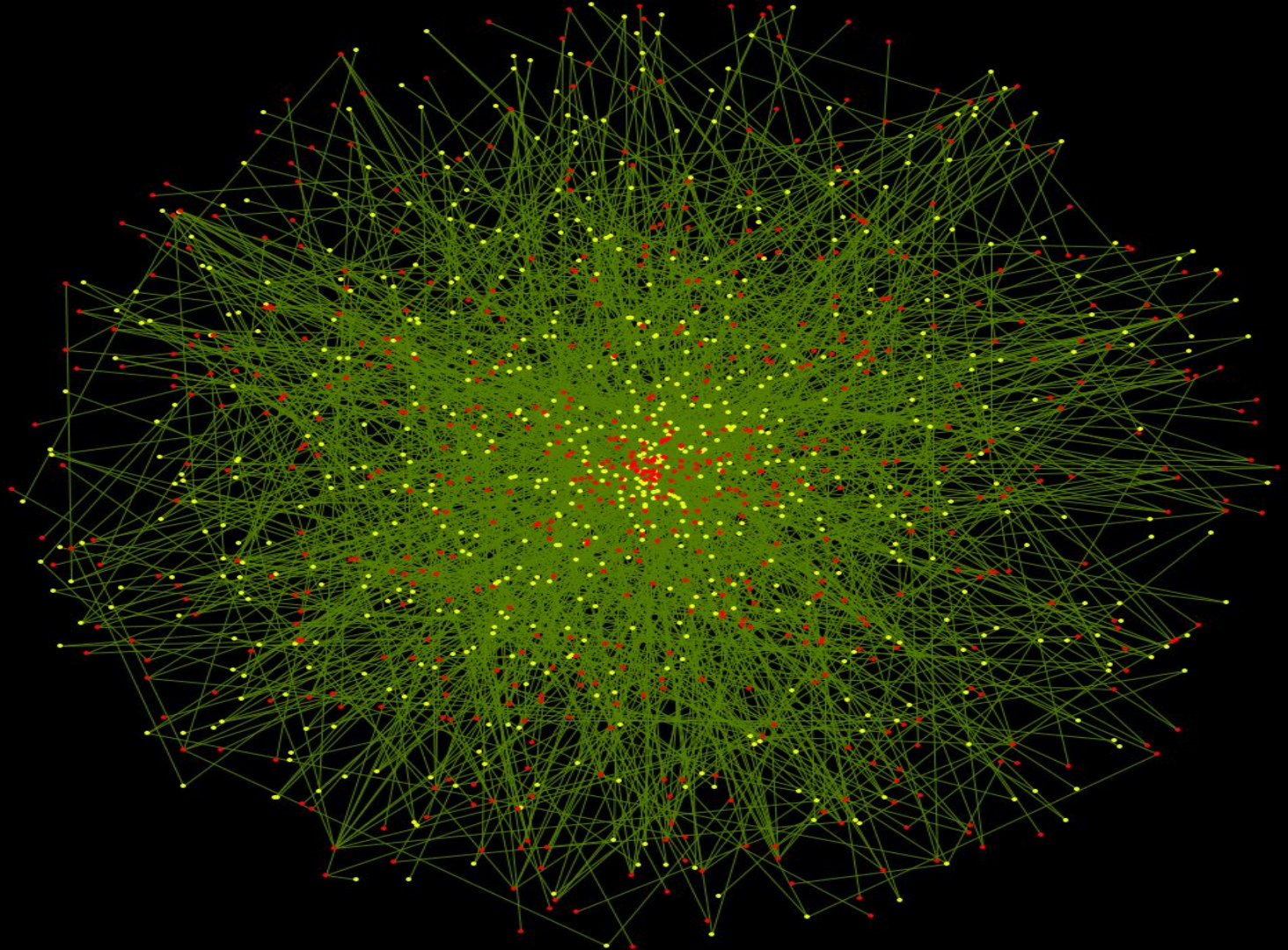


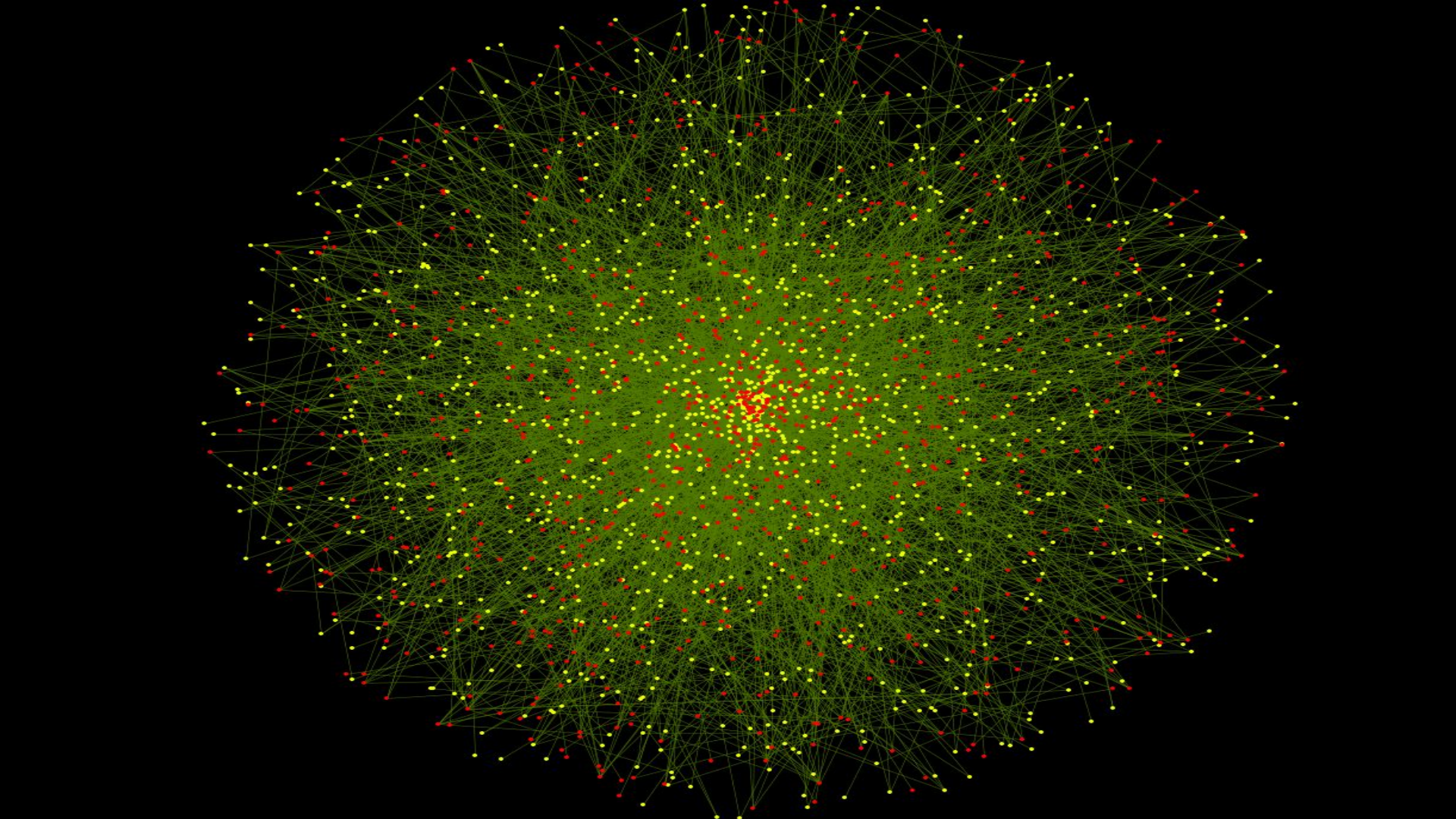


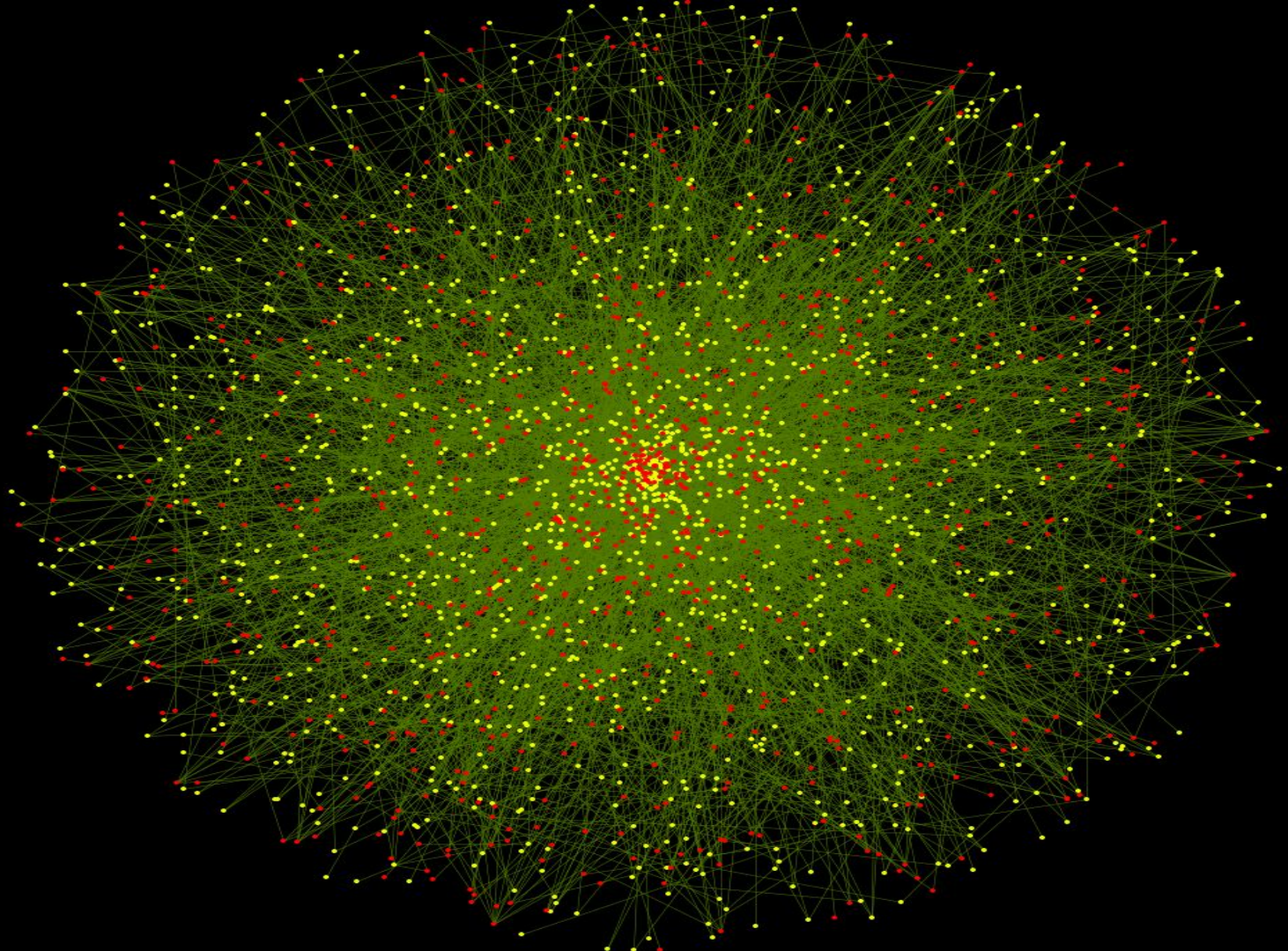


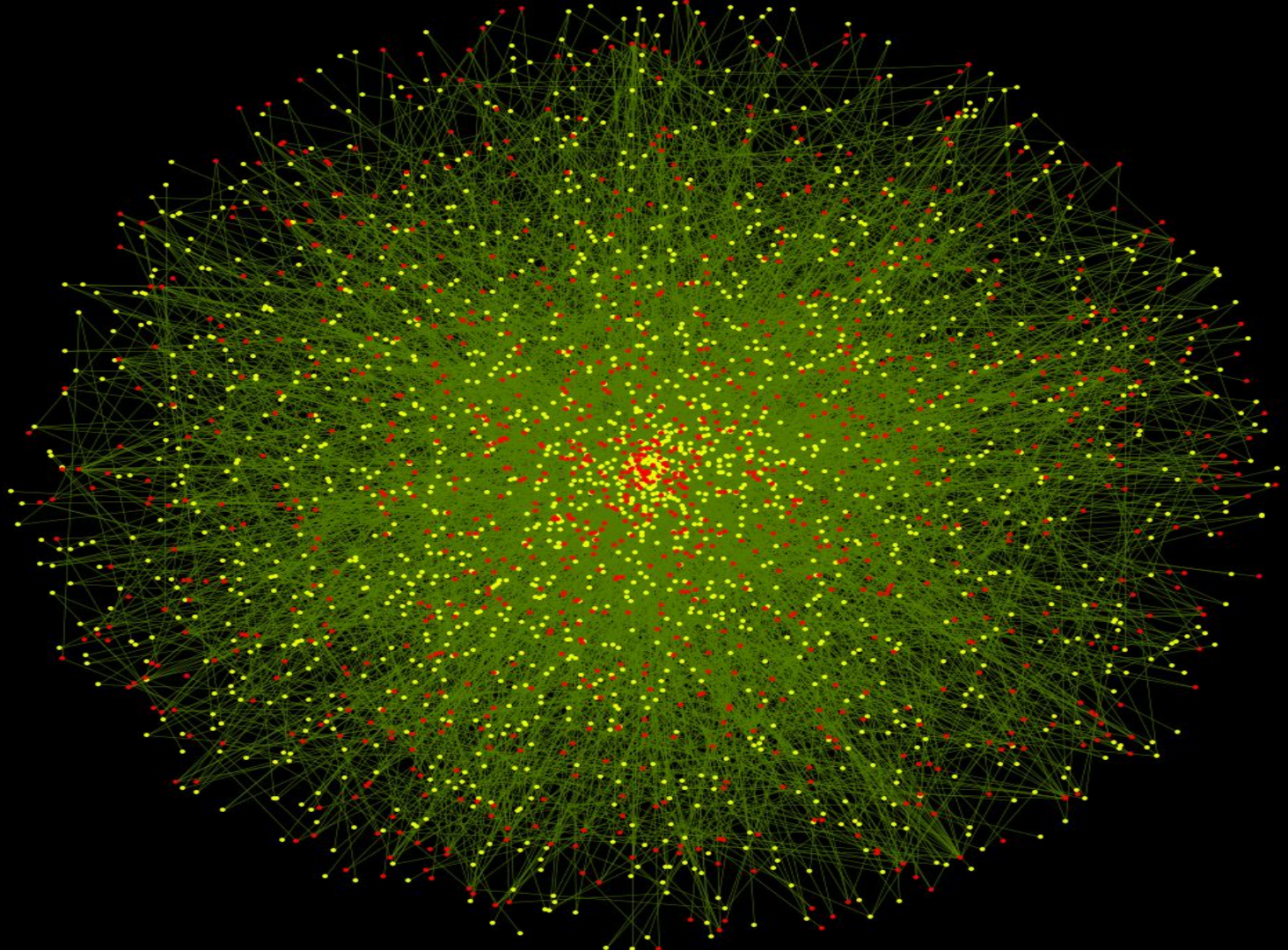


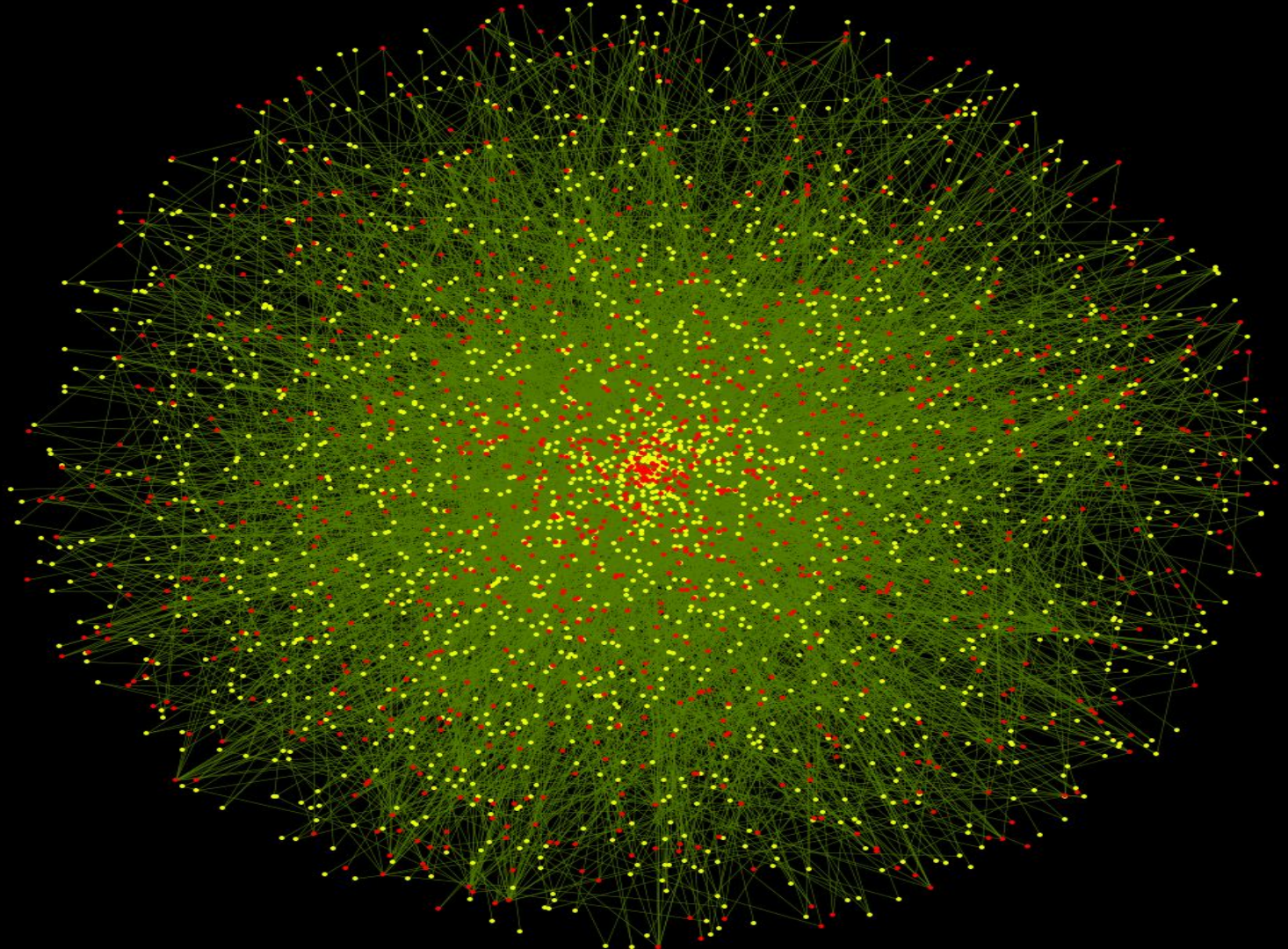


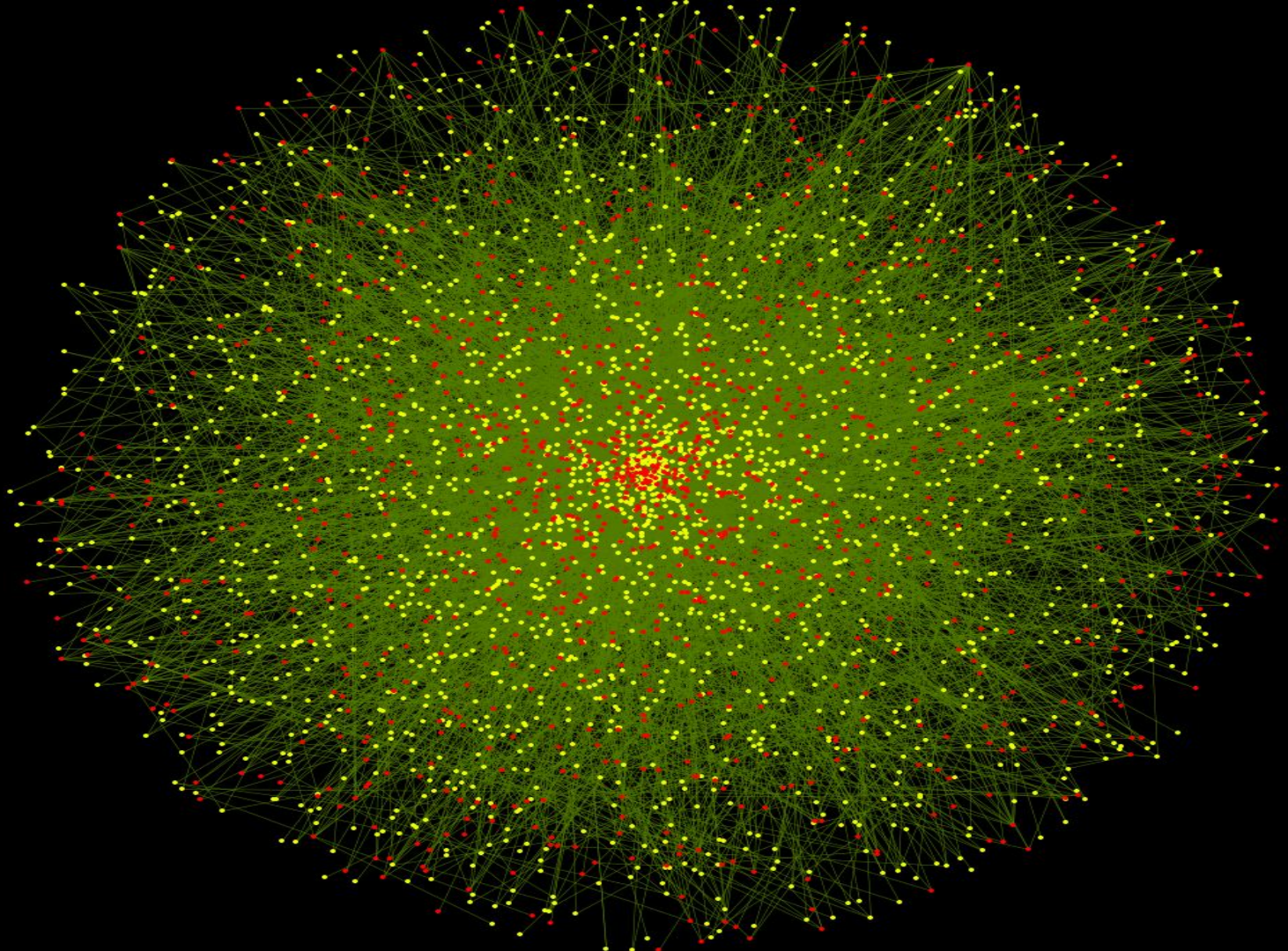


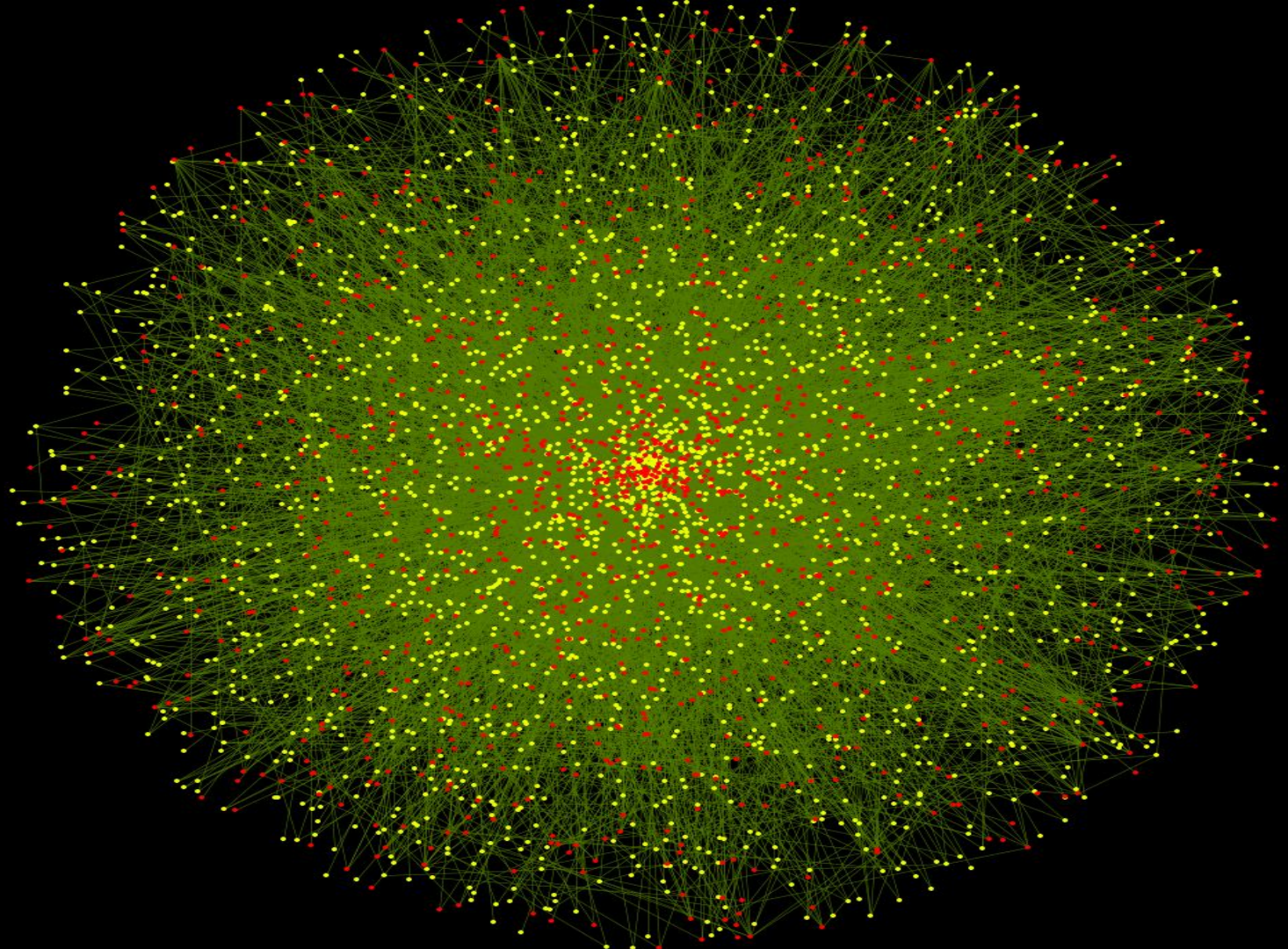


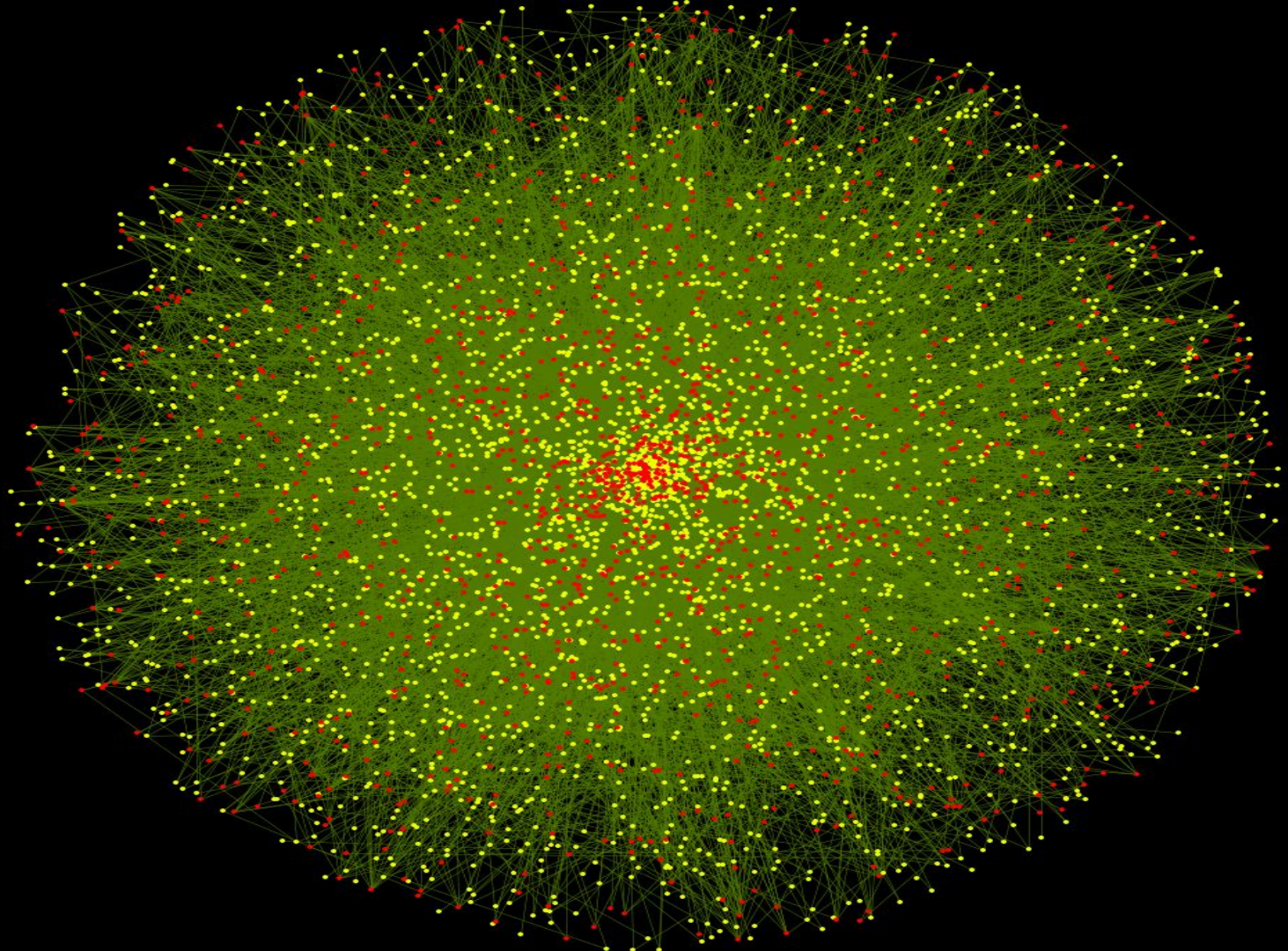


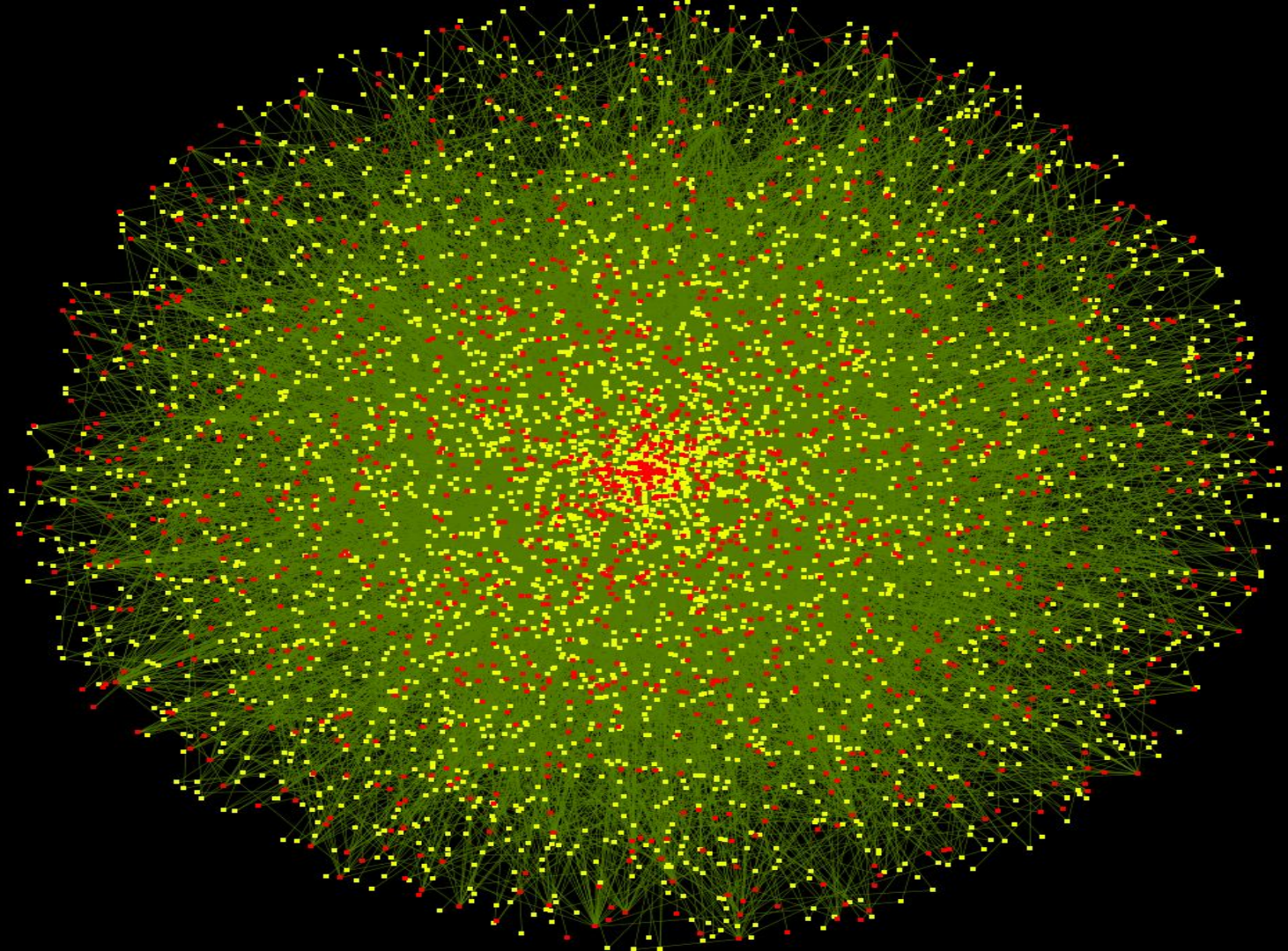


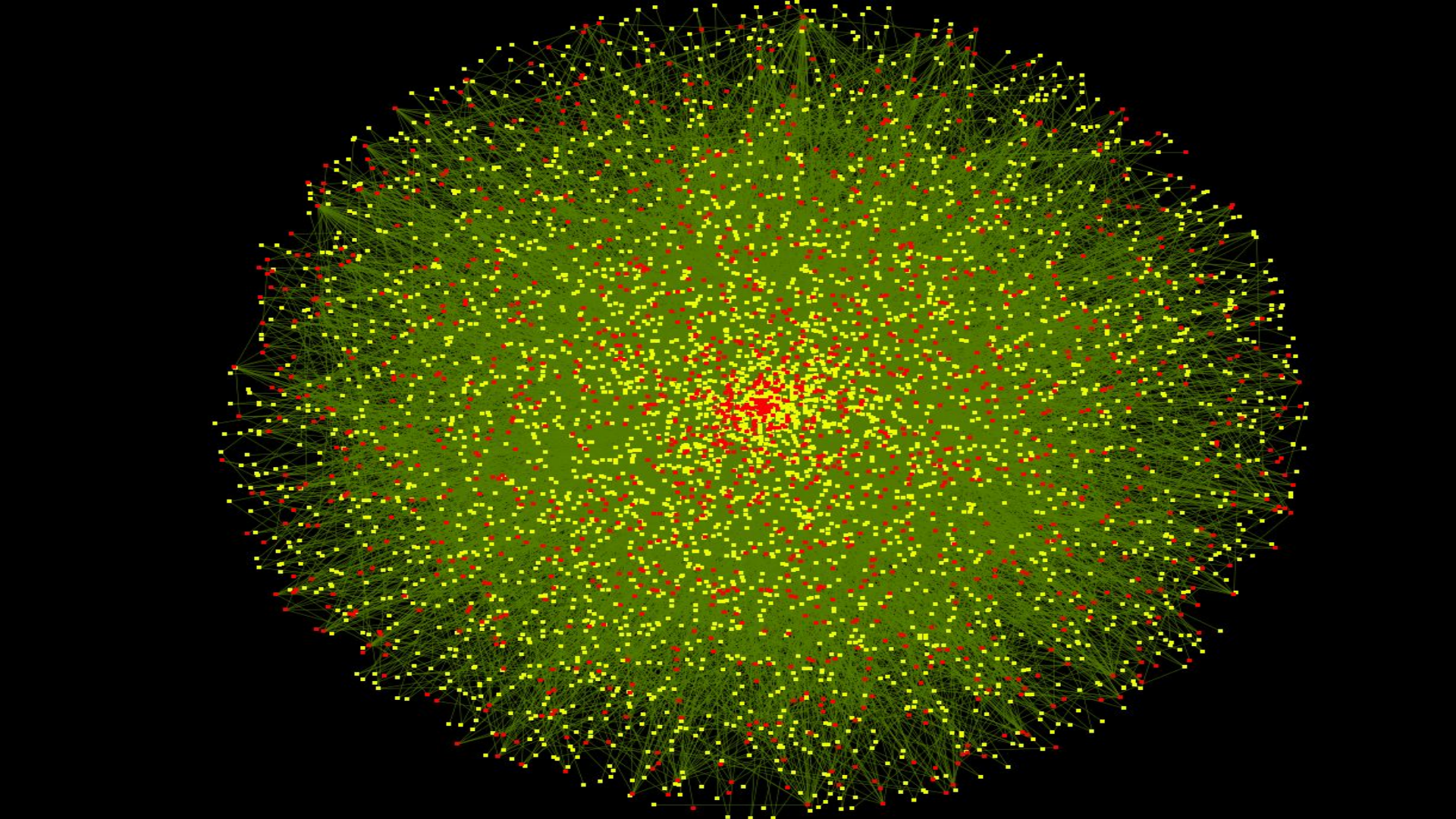


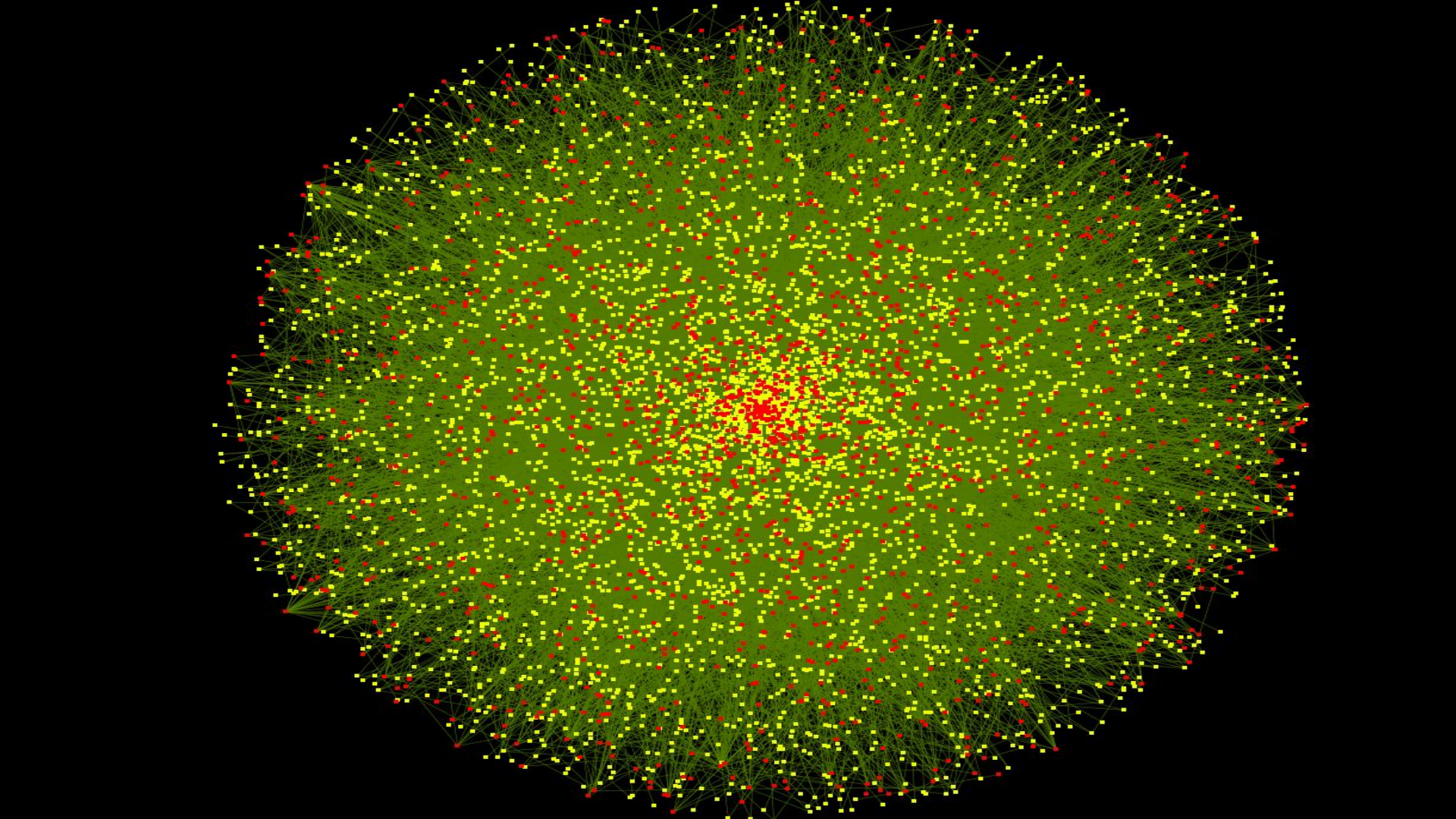








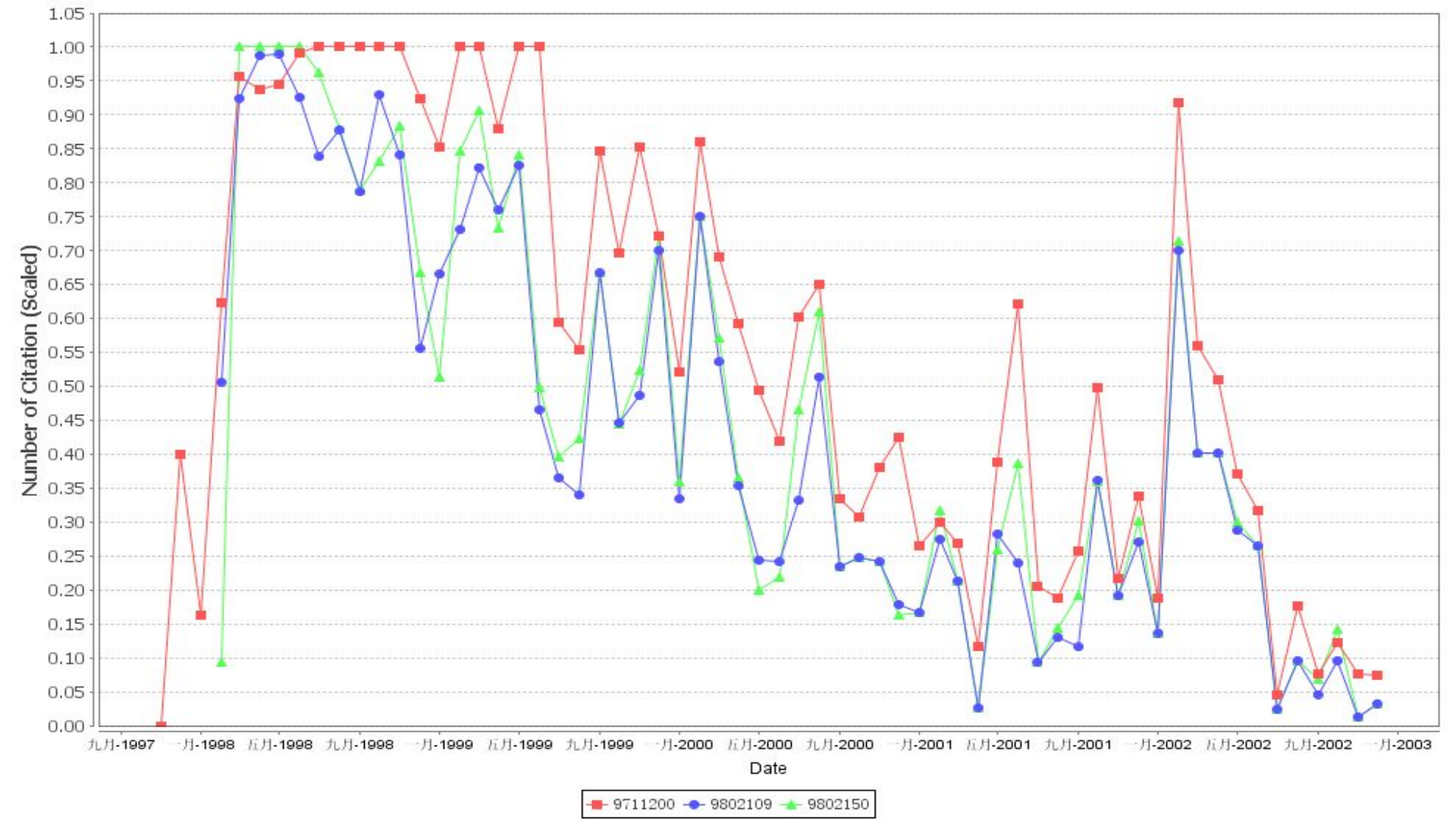


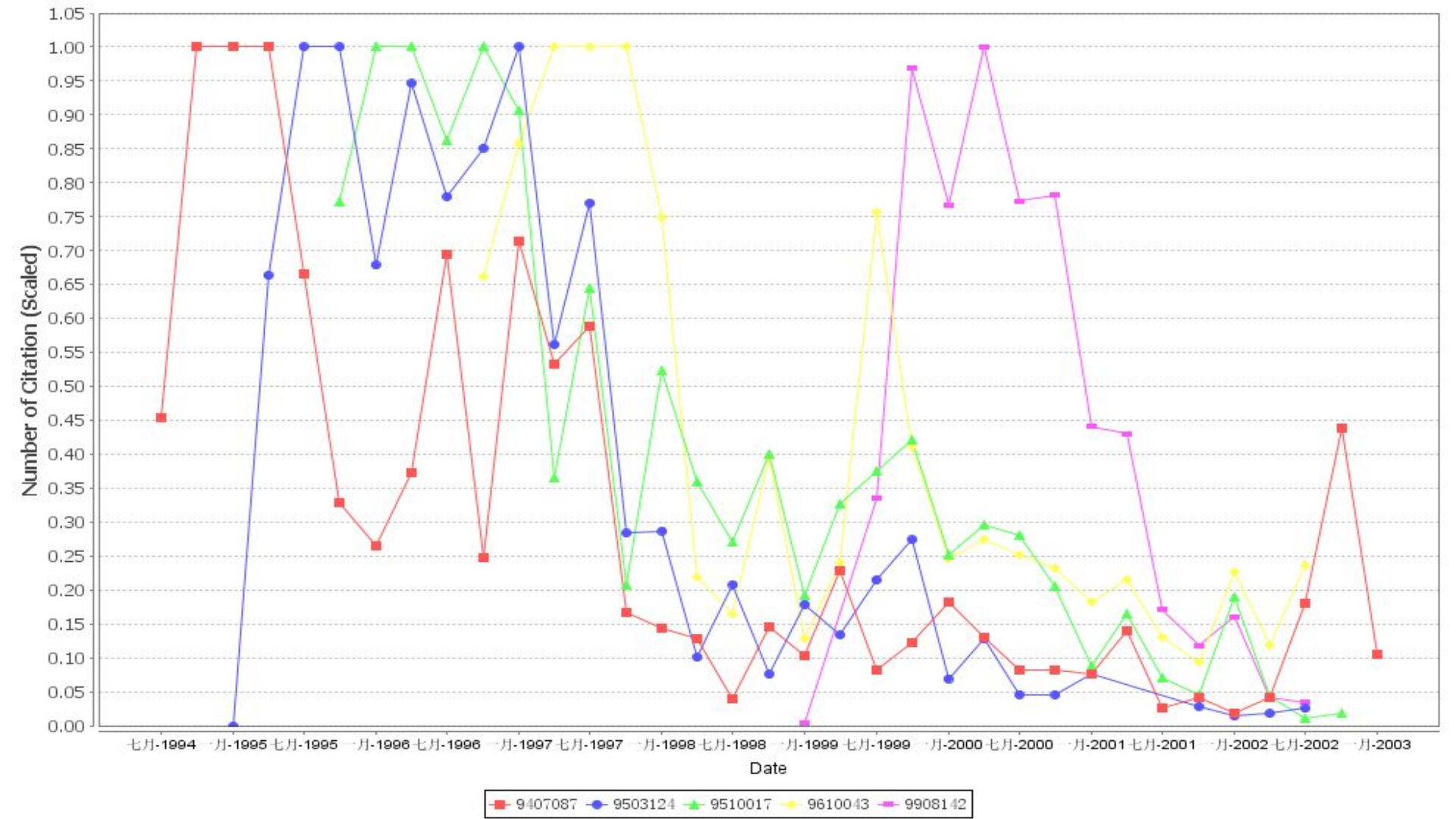


Paper Attributes Analysis

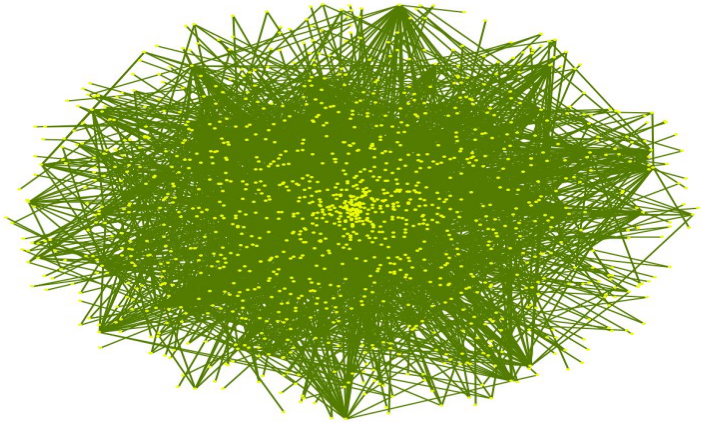
Rank	Paper ID	Number of Citation
1	9711200	2414
2	9802150	1775
3	9802109	1641
4	9407087	1299
5	9610043	1199
6	9510017	1155
7	9908142	1144
8	9503124	1114

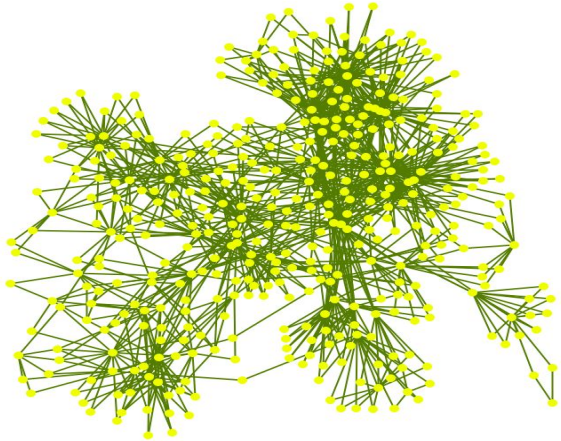
Dynamic Analysis


Number of Citation (Scaled)
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
九月-1997
一月-1998
五月-1998
九月-1998
一月-1999
五月-1999
九月-1999
一月-2000
五月-2000
九月-2000
一月-2001
五月-2001
九月-2001
一月-2002
五月-2002
九月-2002
一月-2003
Date
9711200
9802109
9802150


Number of Citation (Scaled)
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
七月-1994
月-1995
七月-1995
月-1996
七月-1996
月-1997
七月-1997
月-1998
七月-1998
月-1999
七月-1999
月-2000
七月-2000
月-2001
七月-2001
月-2002
七月-2002
月-2003
Date
9407087
9503124
9510017
9610043
9908142

Paper to Paper Analysis





2000/02 Entire Network

2000/02 Core Network

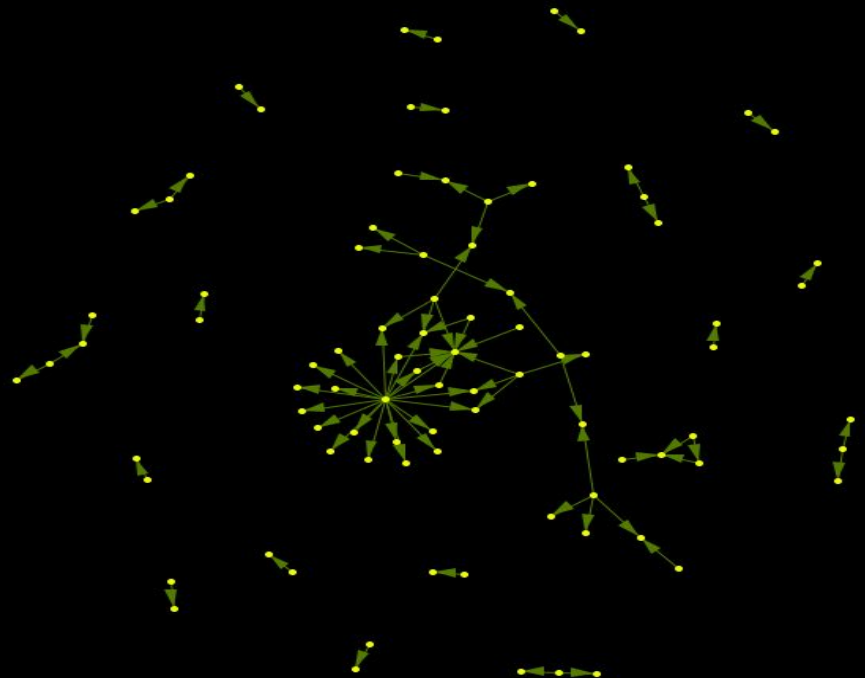
	Entire Network	Core Network
Number of entities	1258	424

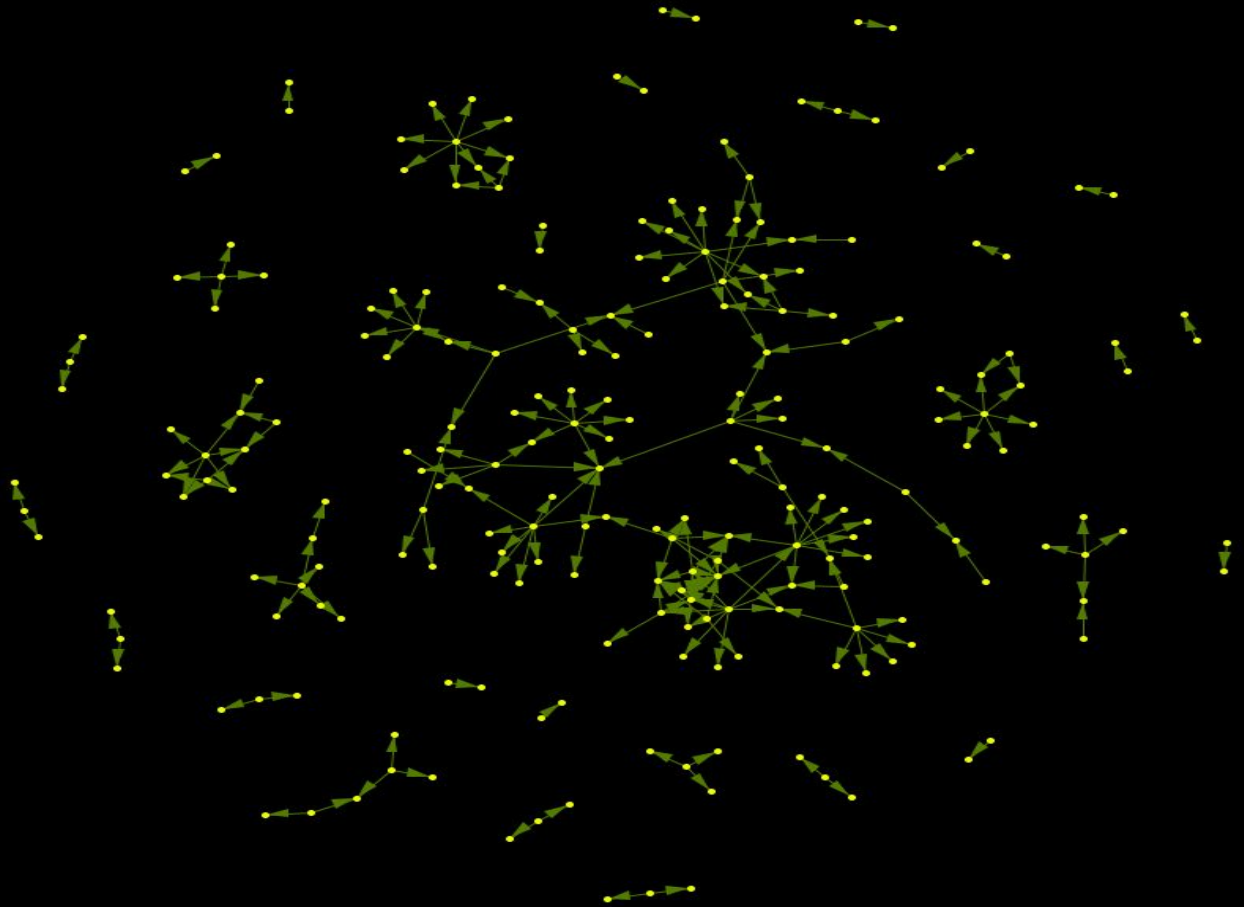
Paper to Paper Analysis

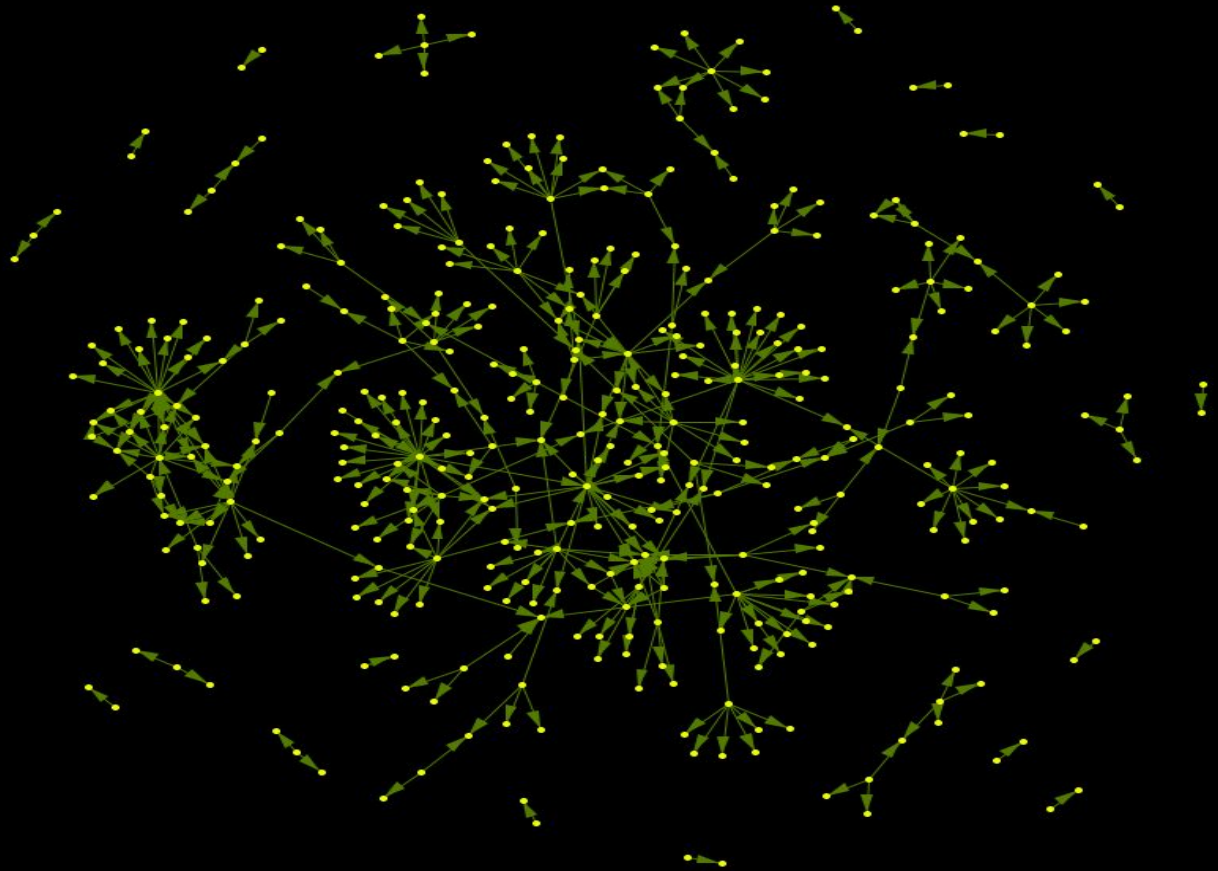
Potentially Influential (betweenness centrality)

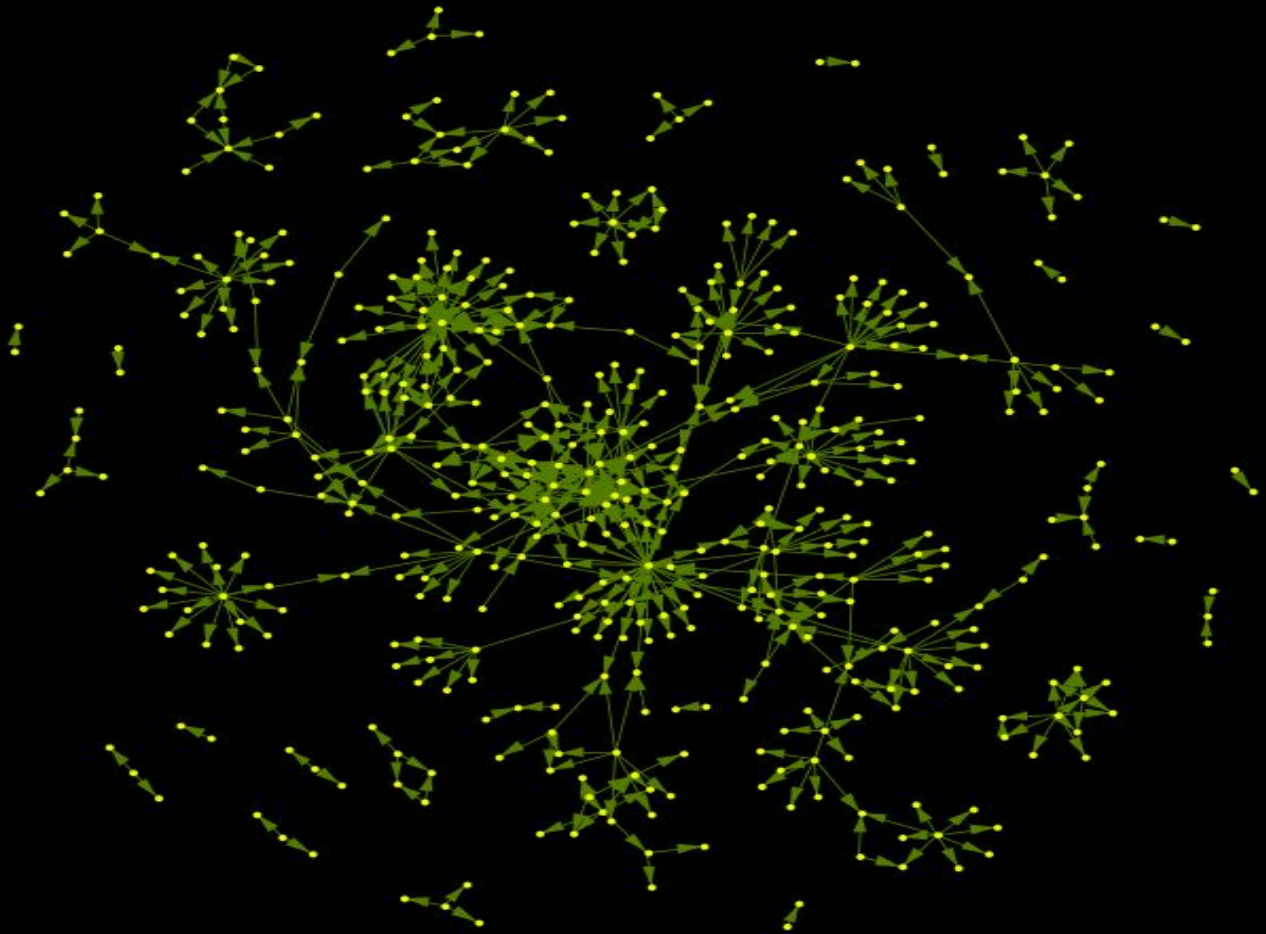
The Betweenness Centrality of node v in a network is defined as: across all node pairs that have a shortest path containing v.

Rank	Paper ID
1	200001033
2	200001002
3	200001206
4	200002160
5	200002161









GeoSpatial Networks

Who was Where When?

- **Remote Capabilities Assessment**
  - Umbrella project sponsored by DTRA
  - Defense Threat Reduction Agency

- **Overarching Goals:**
  - Remotely assess CBRNE capabilities of countries
    - Chemical, biological, nuclear weapons of mass destruction
  - Identify key people/organization involved at each region
  - Trends analysis
    - Early detection of abnormal signals
  - Spatio-temporal trend analysis

Nuclear Research Dataset

- Nuclear Science
  - Political and national security issues
  - Mostly interested in getting country level information

- Roughly 20 years of data, 1992-2010
  - arXiv Cornell pre-print library (Publicly available).
  - Experimental / Theoretical Nuclear Publications

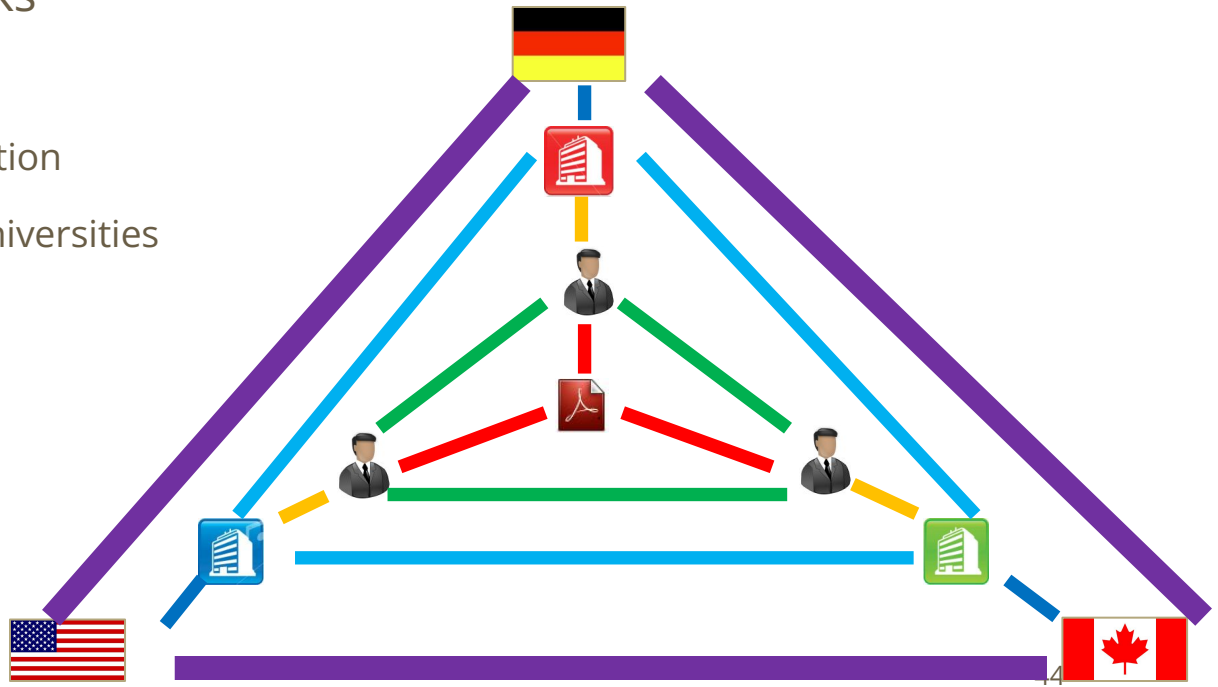
- Construct social networks of nuclear researchers
- Perform spatio-temporal mining & social network analysis

Multi-Level Nuclear Collaboration Networks

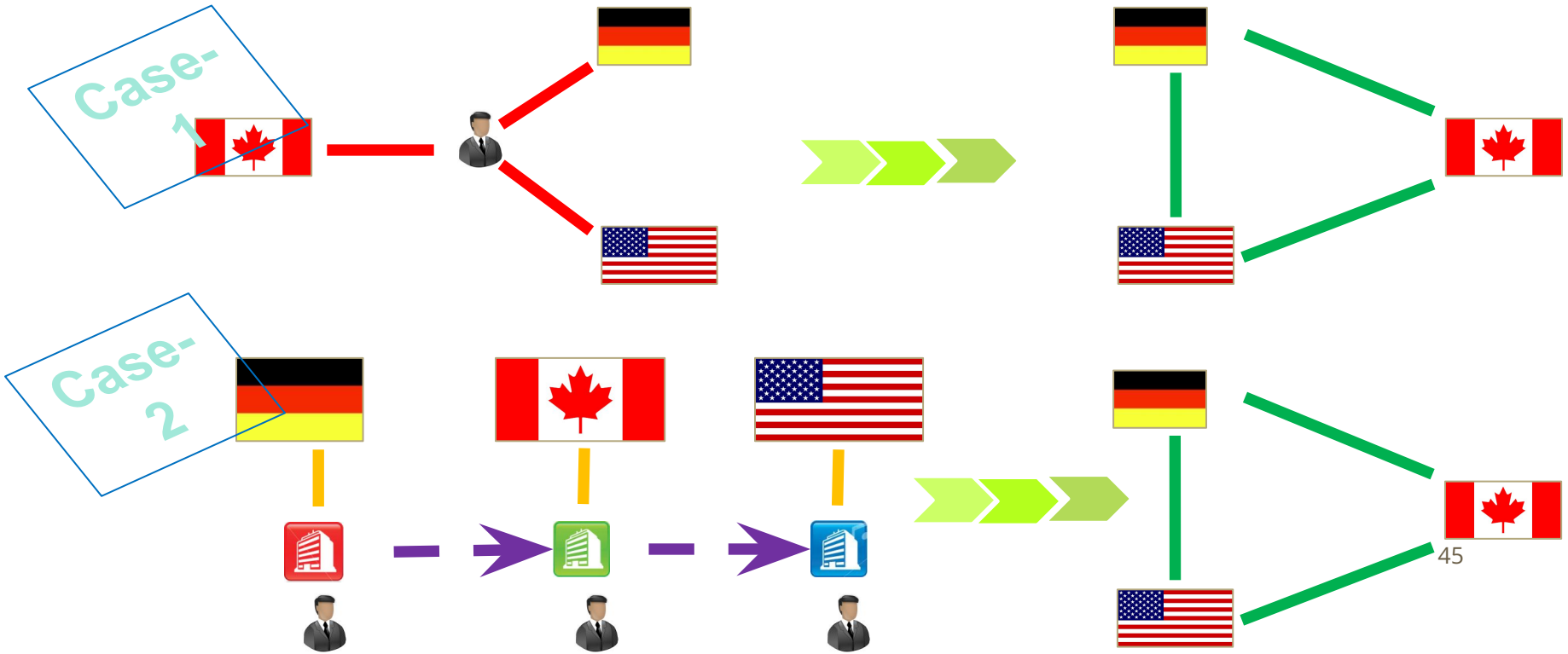
- Country-Level Collaboration Networks
  - Shared Papers
  - Shared Authors

Collaboration Networks (Shared Papers)

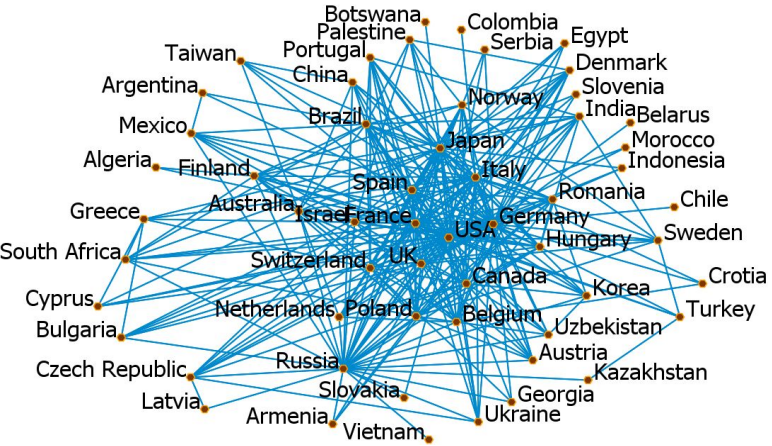
- Spatiotemporal networks
  - Coauthors are linked
  - Each author has an affiliation
  - Collaboration between universities and countries
  - Based on ***Shared Papers***



Collaboration Networks (Shared Authors)

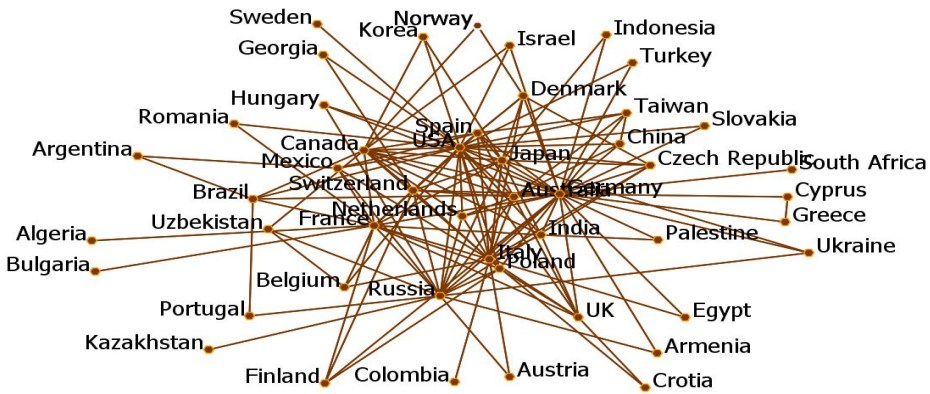


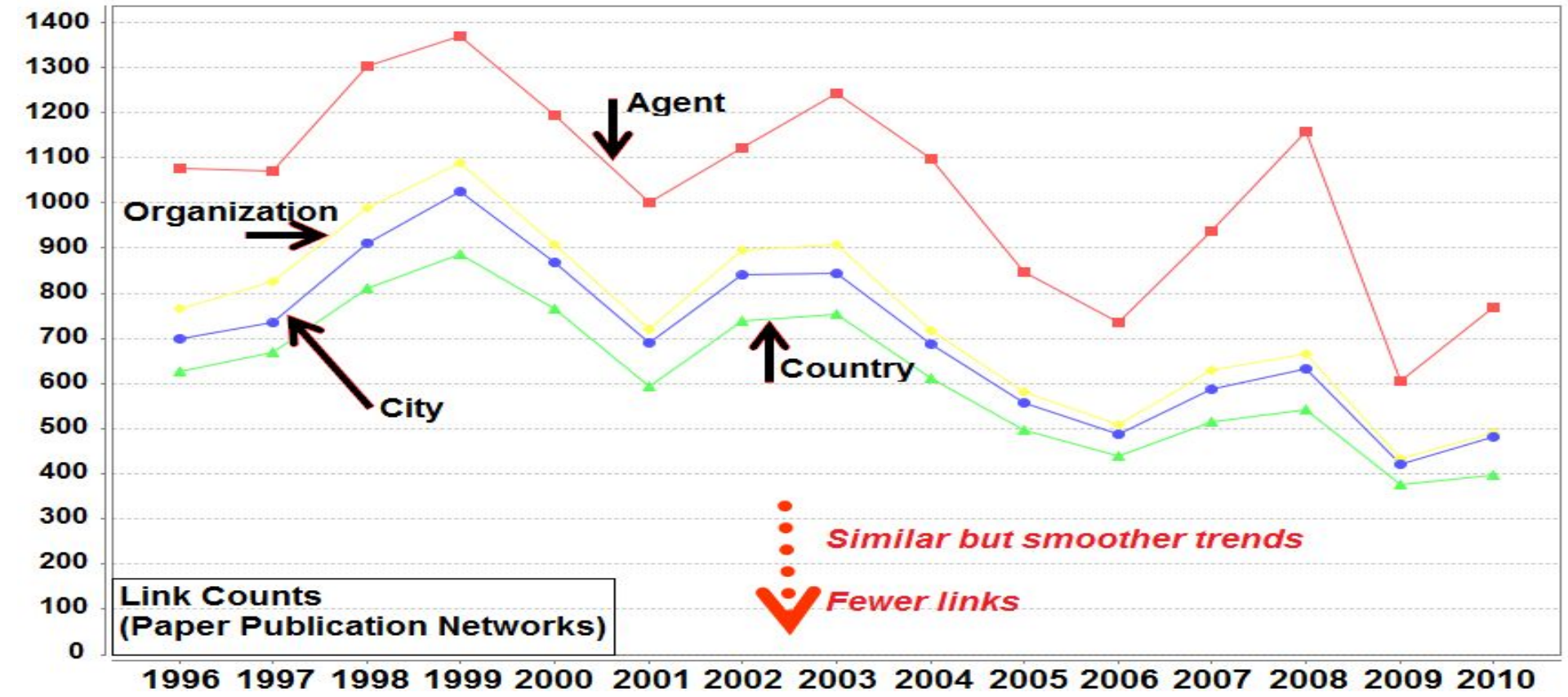
Countries share more Papers than Authors



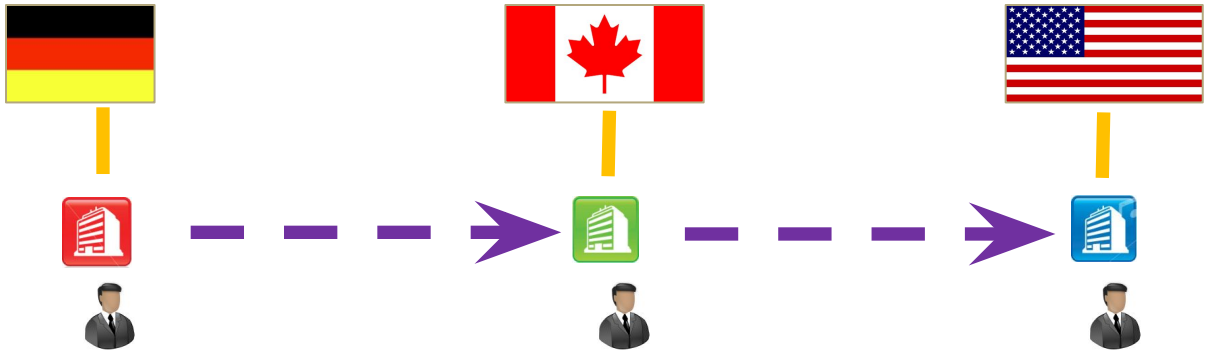
Aggregate Network for Shared Papers

Aggregate Network for Shared Authors




1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
Agent
Organization
City
Country
Similar but smoother trends
Fewer links
Link Counts
(Paper Publication Networks)
1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

Spatiotemporal Mobility Patterns



Relocating Researchers

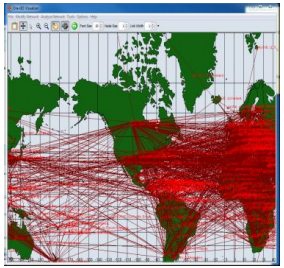
2 pieces of information:

1. Individual Trails
2. Aggregate Transition Matrix

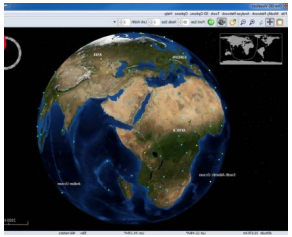
Spatiotemporal Analysis Software

- *ORA 
  - Dynamic meta network development & assessment tool
  - Developed by CASOS Research group @ Carnegie Mellon

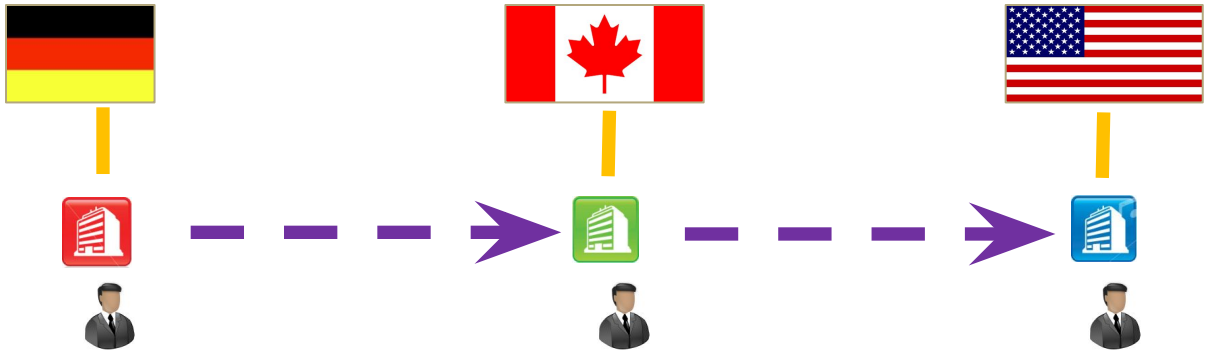
- 2-dimensional map Powered by Openmap



- 3-dimensional map Powered by NASA's WorldWind Java



Back to Mobility Patterns...



Relocating Researchers

2 pieces of information:

1. Individual Trails
2. Aggregate Transition Matrix

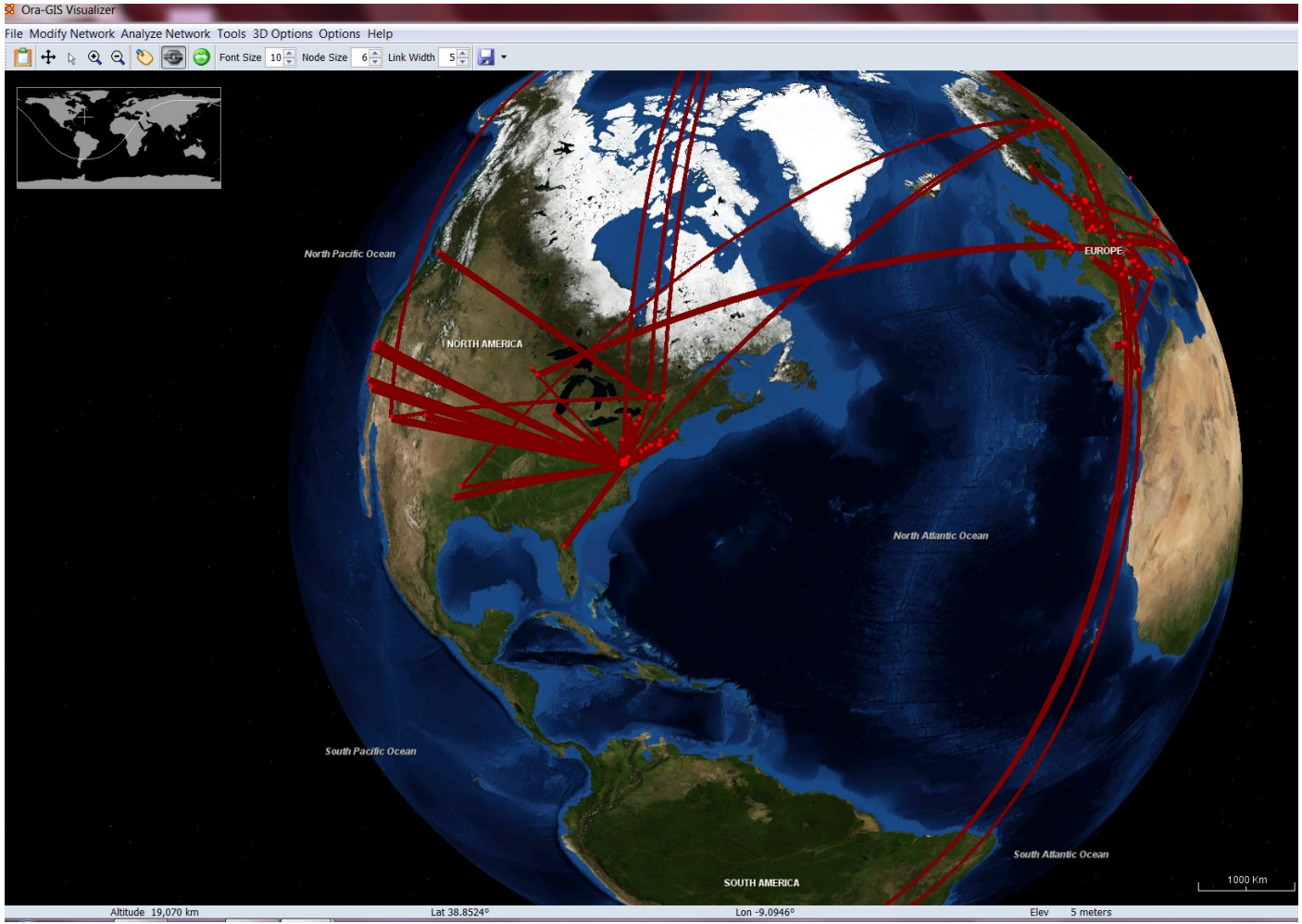
Trails Analysis (Individual Mobility)

- What is **trail**?
  - The paths social agents move in a network

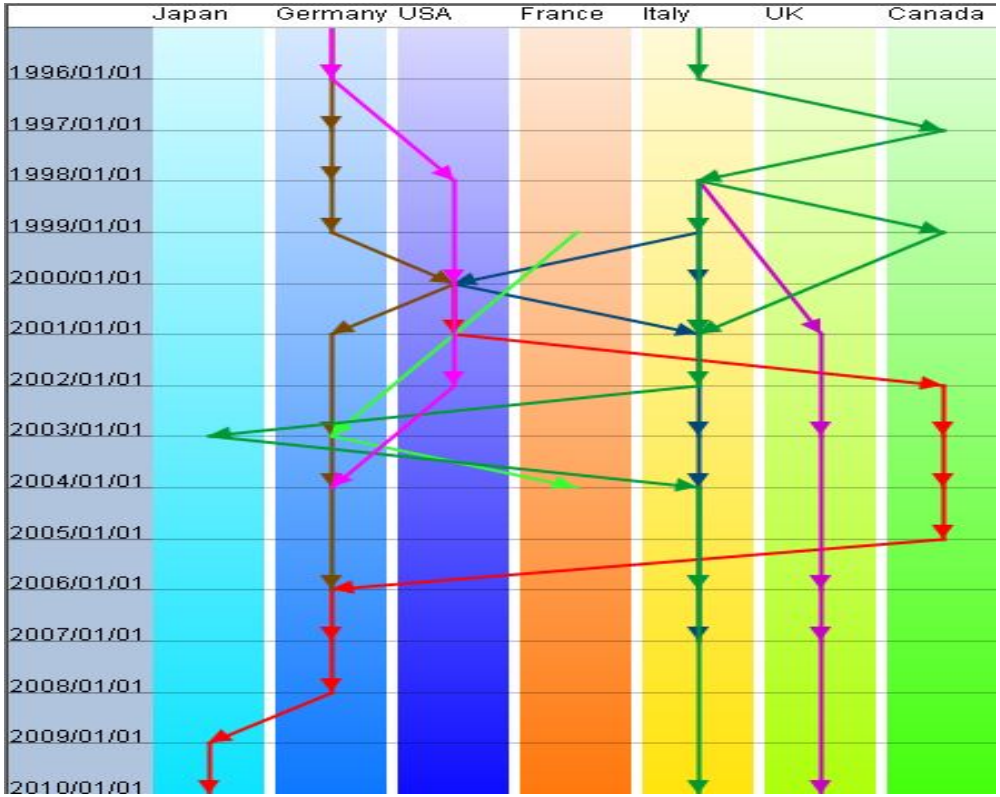
- Track movements of individual actors over time & space.

- Discrete time and space set

- ***Next Slide***: Trails analysis for a number of researchers selected randomly.
  - Condition: The researcher moved at least once.

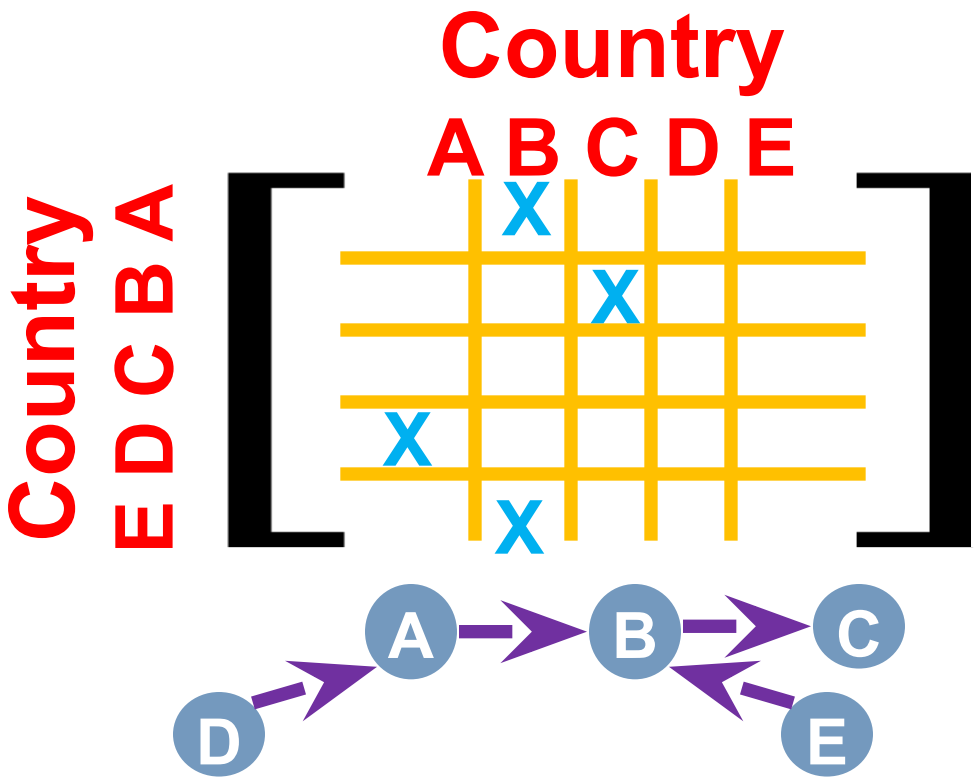

Ora-GIS Visualizer
File Modify Network Analyze Network Tools 3D Options Options Help
Font Size 10 Node Size 6 Link Width 5
North Pacific Ocean
NORTH AMERICA
EUROPE
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
SOUTH AMERICA
1000 Km
Altitude 19,070 km
Lat 38.8524°
Lon -9.0946°
Elev 5 meters

Researcher Mobility Patterns (Trails Analysis)



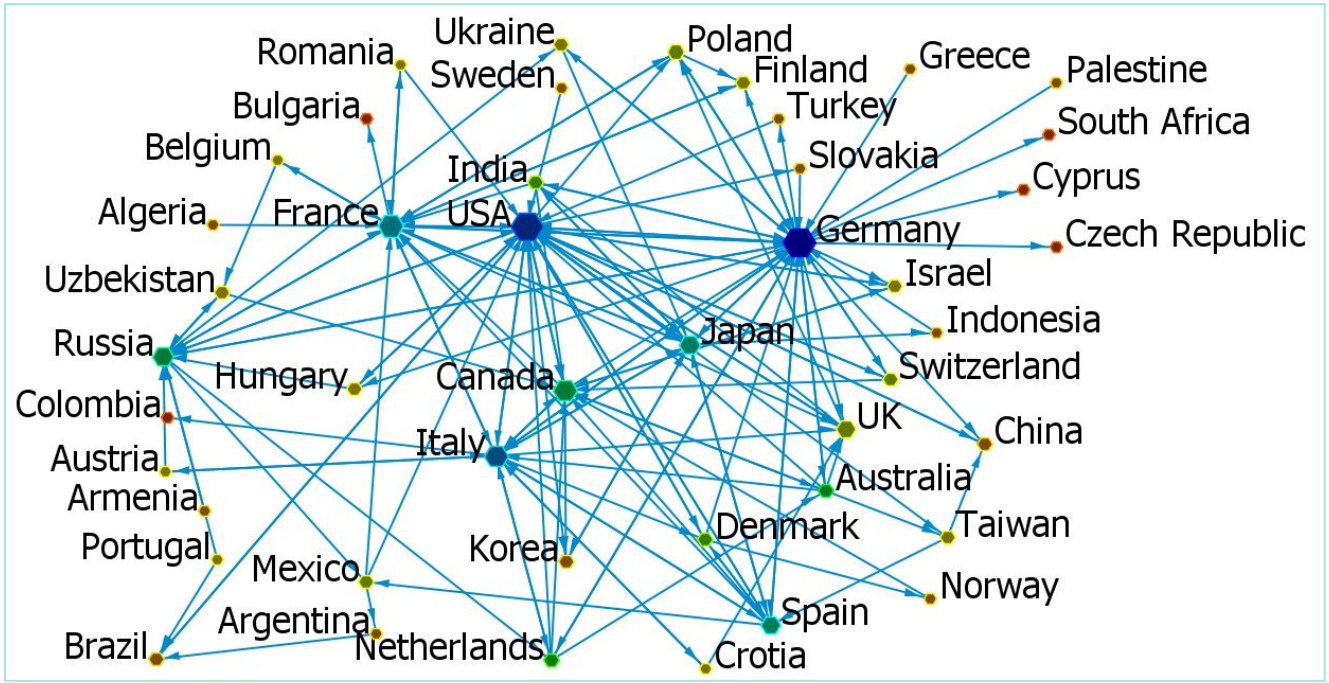
- People returning back to their countries
- Post-Doc research is common

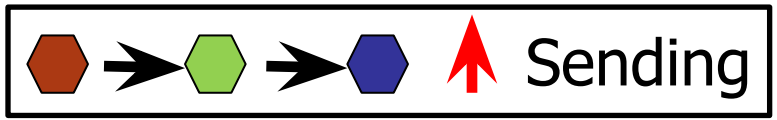
Aggregate Transition Matrix (Overall Mobility)

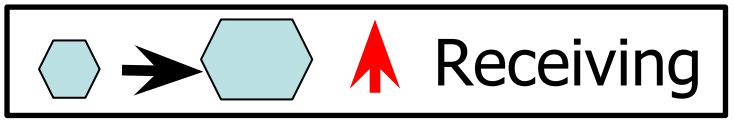


- Country x Country Matrix
- Directed link from **A** --> **B** if a researcher traveled from **A** to **B** directly.
- Extract information for *all* researchers across *all* papers.

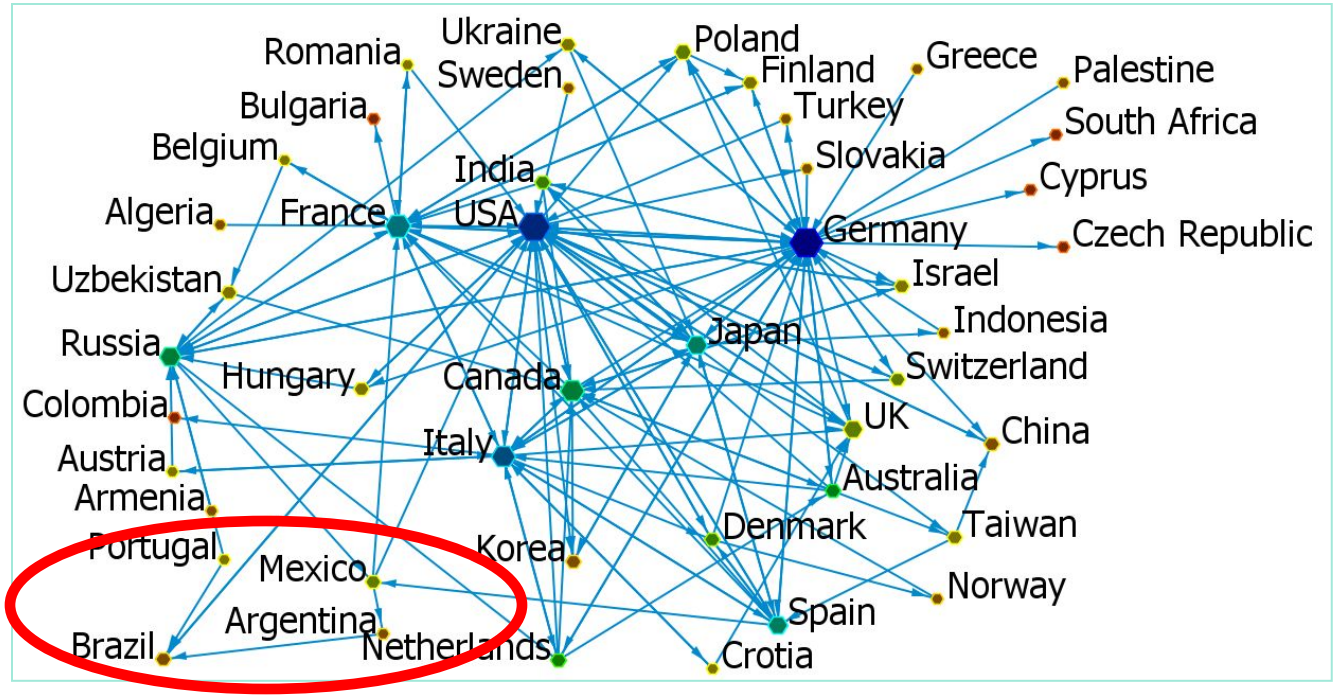
How Information Travels? (Researcher Mobility)

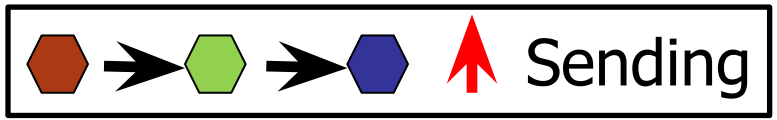


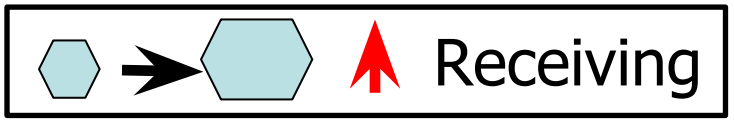




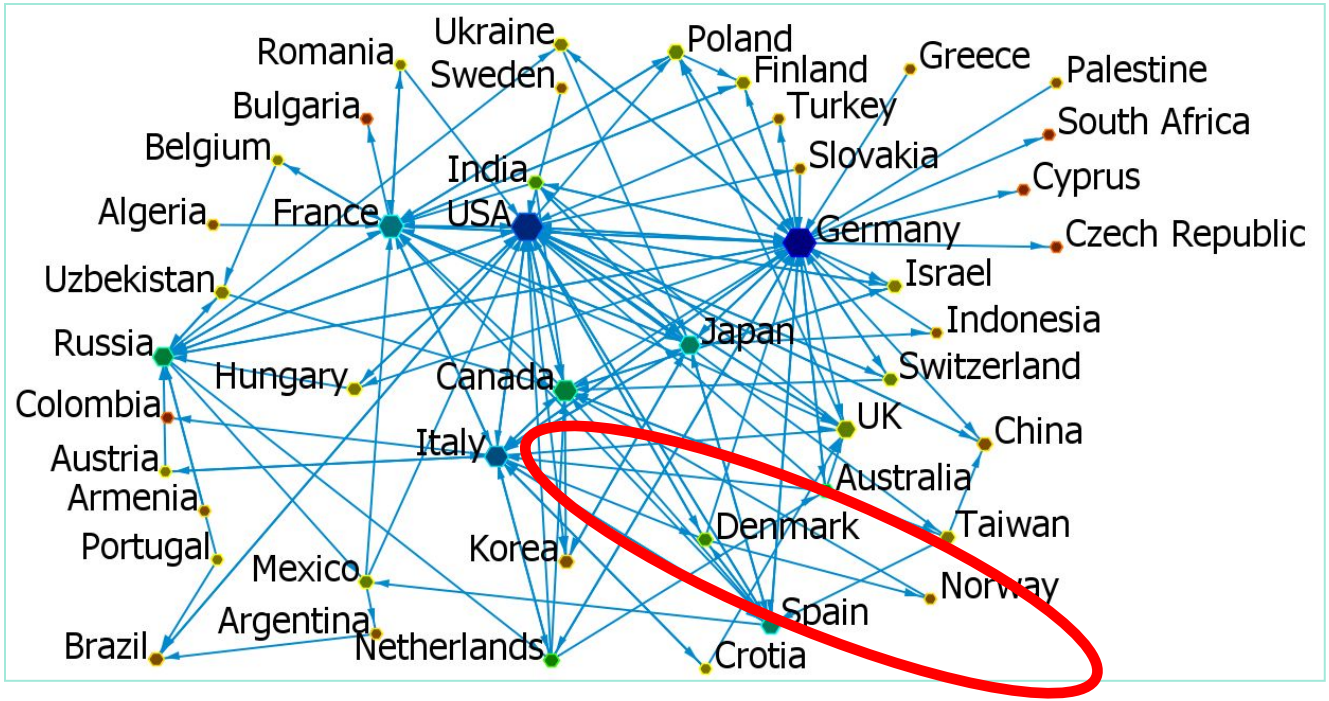
How Information Travels? (Researcher Mobility)

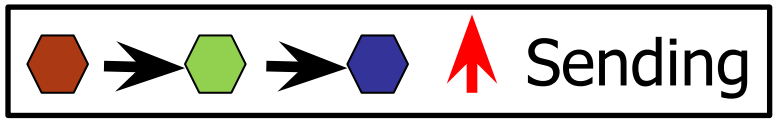


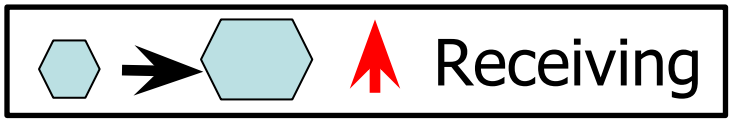




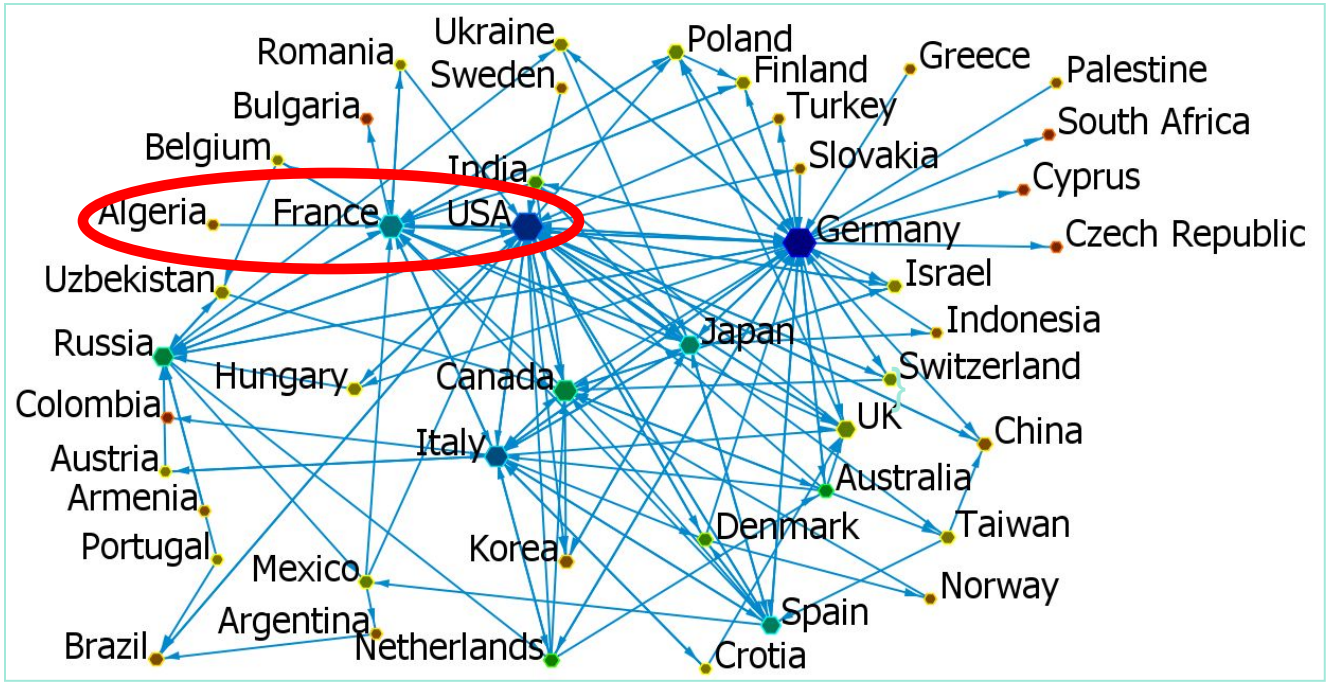
How Information Travels? (Researcher Mobility)

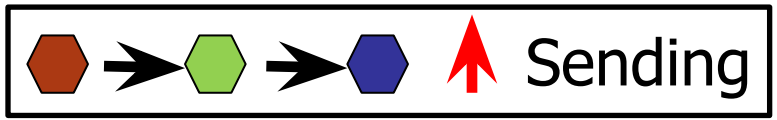


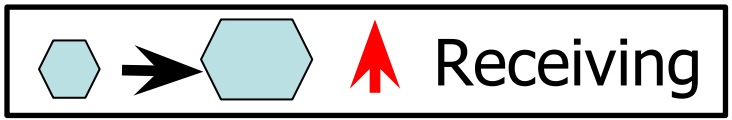




How Information Travels? (Researcher Mobility)







Country Level Centralities

- Various centrality metrics from social network analysis
  - **Degree Centrality:** Most Connections
  - **Betweenness Centrality:** On most shortest paths. Who are influential intermediates/brokers?
  - **EigenVector Centrality:** Cluster leader, who has the power to mobilize others?

Degree	Betweenness	EigenVector
USA	Germany	USA
Germany	Russia	Germany
France	USA	France
Russia	Italy	Canada
Italy	Mexico	Russia
Japan	Canada	Japan

Recap: *Patterns to Note*

- Highest senders are also the highest receivers.
  - Due to many people going back to their own countries.
- Postdoc is very common for physics.
- **Researcher Exchange Patterns:**
  - Speaking the same/similar language.
    - E.g., Brazil, Argentina, Mexico, and Portugal, Spain exchange researchers. They all speak Spanish or a language with some mutual intelligibility.
  - Geographically proximity.
    - E.g. Croatia/Italy.
  - Historical connections
    - Algeria was a French colony.

AARHUS UNIVERSITET

UNIVERSITY OF WOLVERHAMPTON

Co-Read, Co-Tweet and Co-Citation Networks

Fereshteh Didegah

Danish Centre for Studies in Research & Research Policy (CFA), Aarhus University, Aarhus, Denmark. E-mail: fdid@ps.au.dk

Mike Thelwall

Statistical Cybermetrics Research Group, University of Wolverhampton, Wolverhampton, UK. E-mail: m.thelwall@wlv.ac.uk

Link to paper

3:AM BUCHAREST 2016

Previous research

- Moderate correlation between citation counts and Mendeley readership (Mohammadi & Thelwall, 2013).
- Weak correlation between citation counts and number of tweets (Thelwall *et al*., 2013).

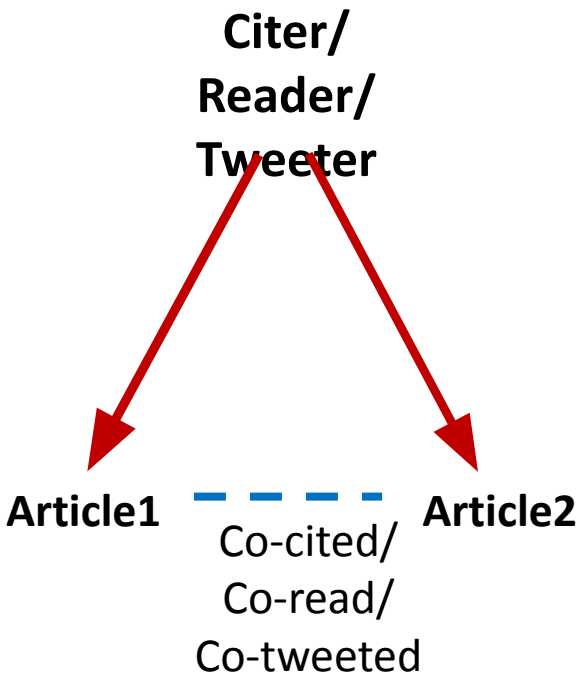
Current research

- The relationship between citations, Mendeley readership and tweets at the network level (User behaviour perspective).

3:AM BUCHAREST 2016

Co-networks

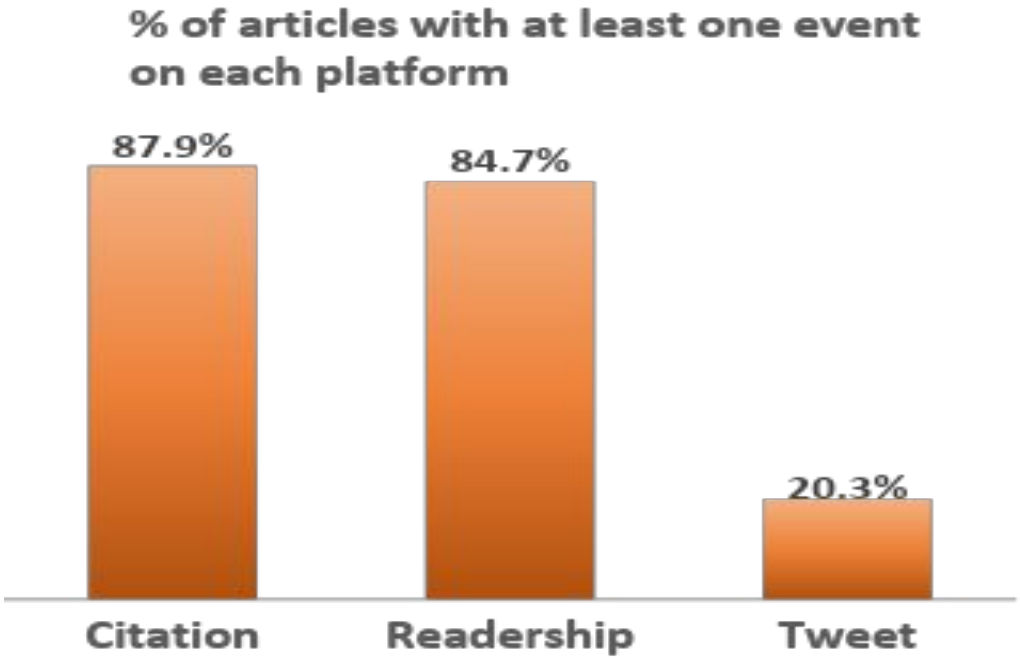
- **Co-citation network:** connects pairs of articles cited by the same citer.
- **Mendeley co-read network:** connects pairs of articles read by the same readers.
- **Co-tweet network:** connects pairs of articles tweeted by the same tweeter.

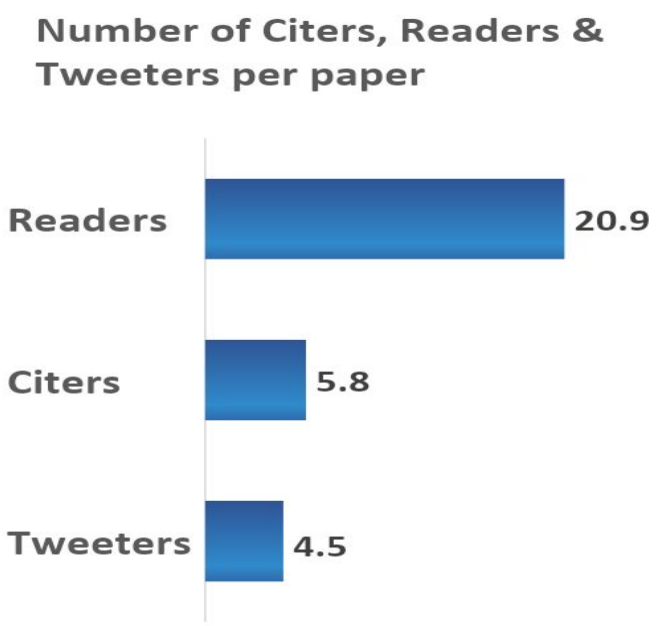


3:AM BUCHAREST 2016

Data

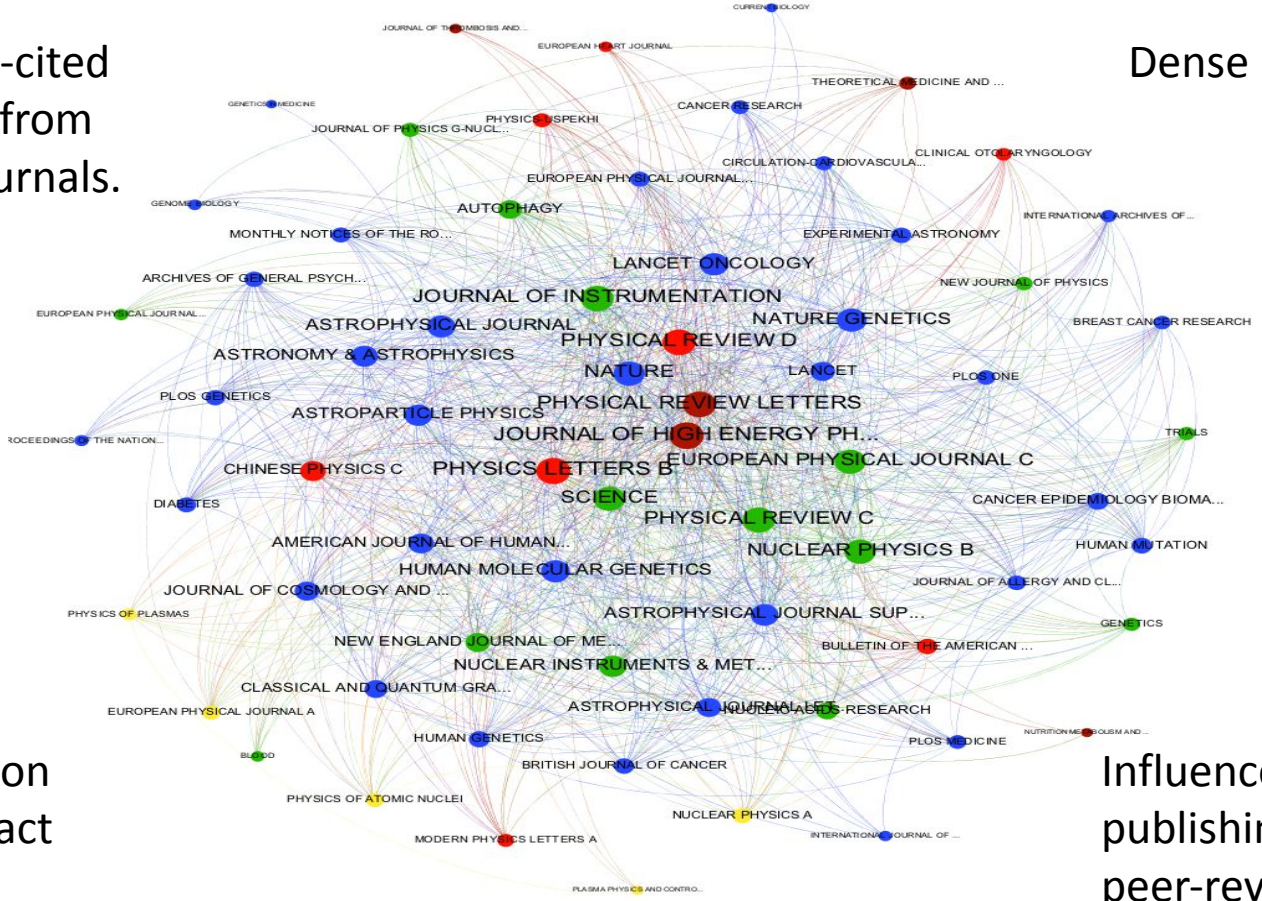
- All articles from **Web of Science** (**WoS**) published in **2012** – a total of 1.1 million articles.
- Citation data from WoS.
- Mendeley readership data from Mendeley.
- Twitter data from altmetric.com.


% of articles with at least one event on each platform
87.9%
84.7%
20.3%
Citation
Readership
Tweet


Number of Citers, Readers & Tweeters per paper
Readers
20.9
Citers
5.8
Tweeters
4.5

Co-cited journals in WoS

85.8% of co-cited articles are from different journals.



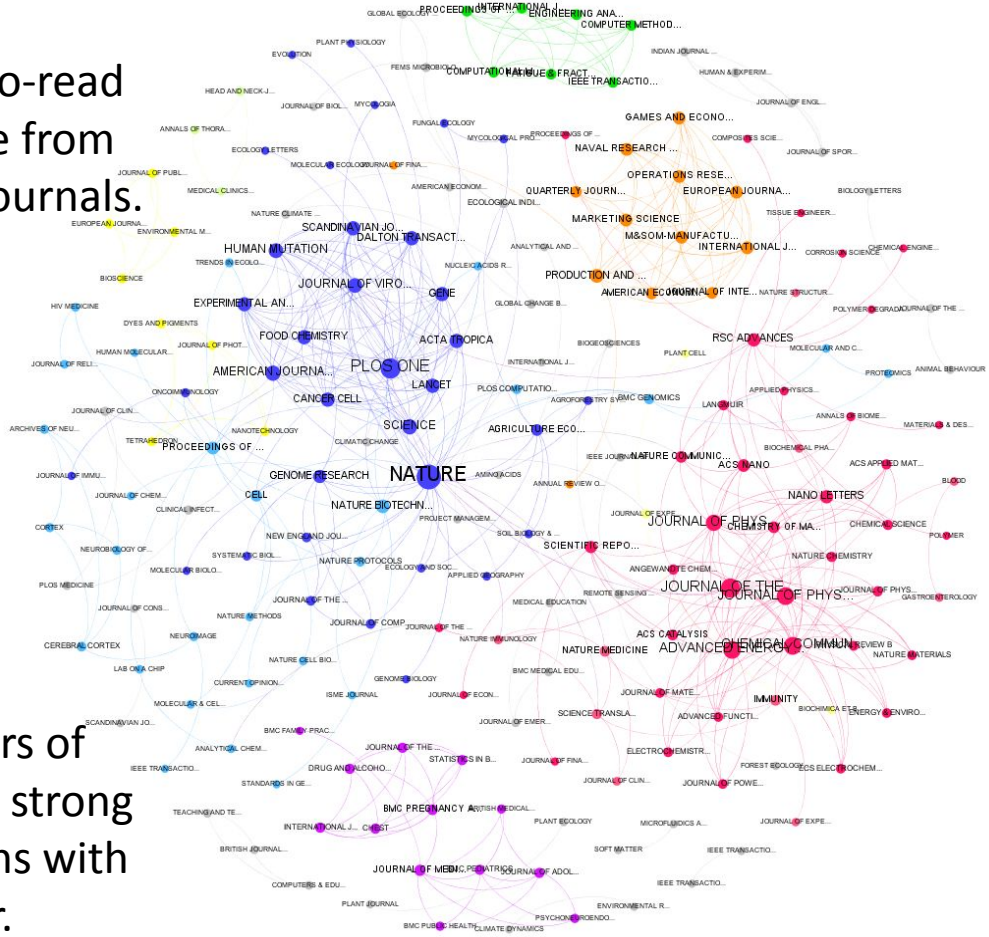
Dense network

Concentration of high impact journals.

Influenced by publishing and peer-review process restrictions

Co-read journals in Mendeley

81.7% of co-read articles are from different journals.



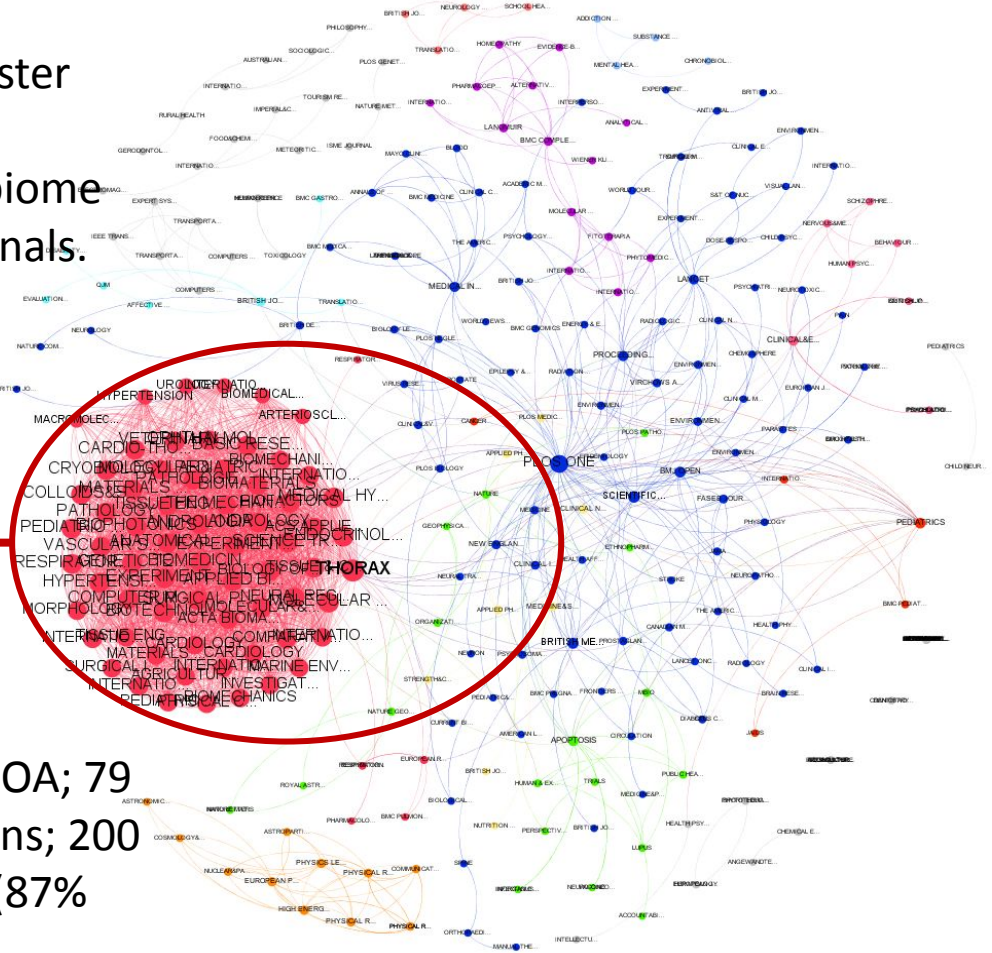
Few clusters of journals in strong connections with each other.

Less dense network

Readers have no restrictions for adding articles to their library.

Co-tweeted journals

Dense cluster of 60 medical/biomedical journals.

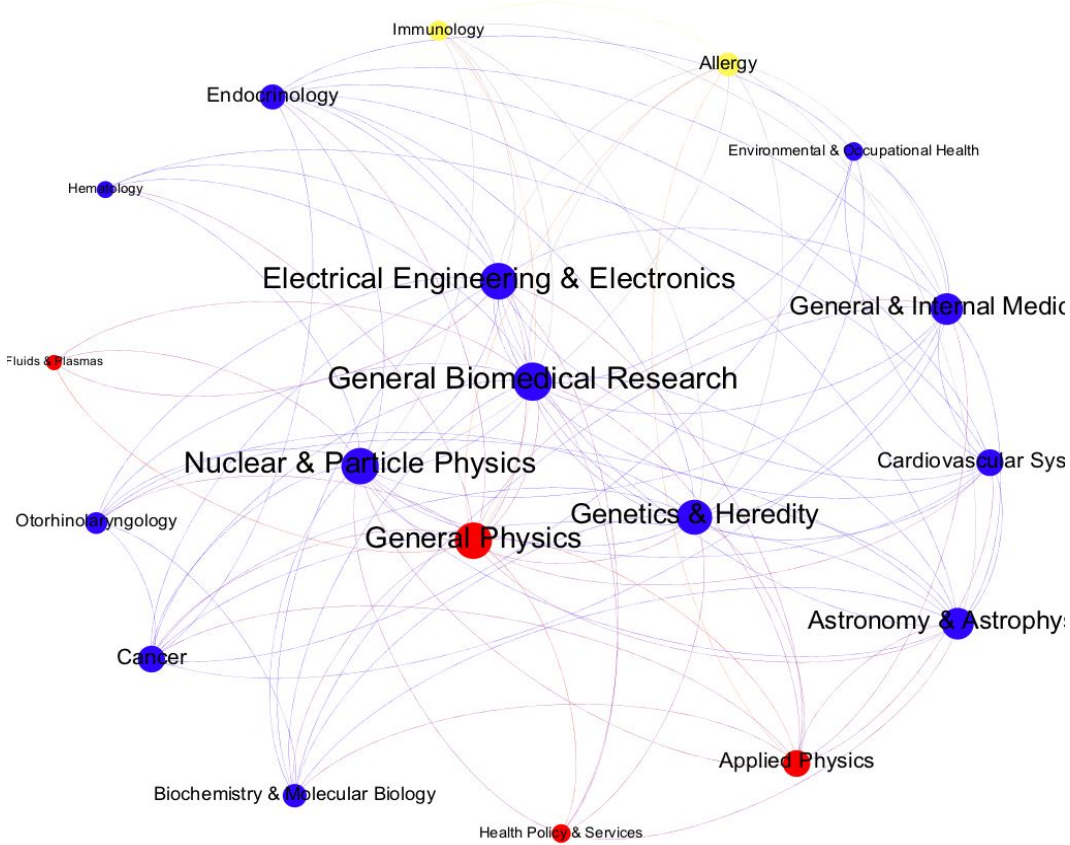


THORAX: OA; 79 connections; 200 tweeters (87% human)

36.5% of co-tweeted articles are from different journals. (**63.5%** the same)

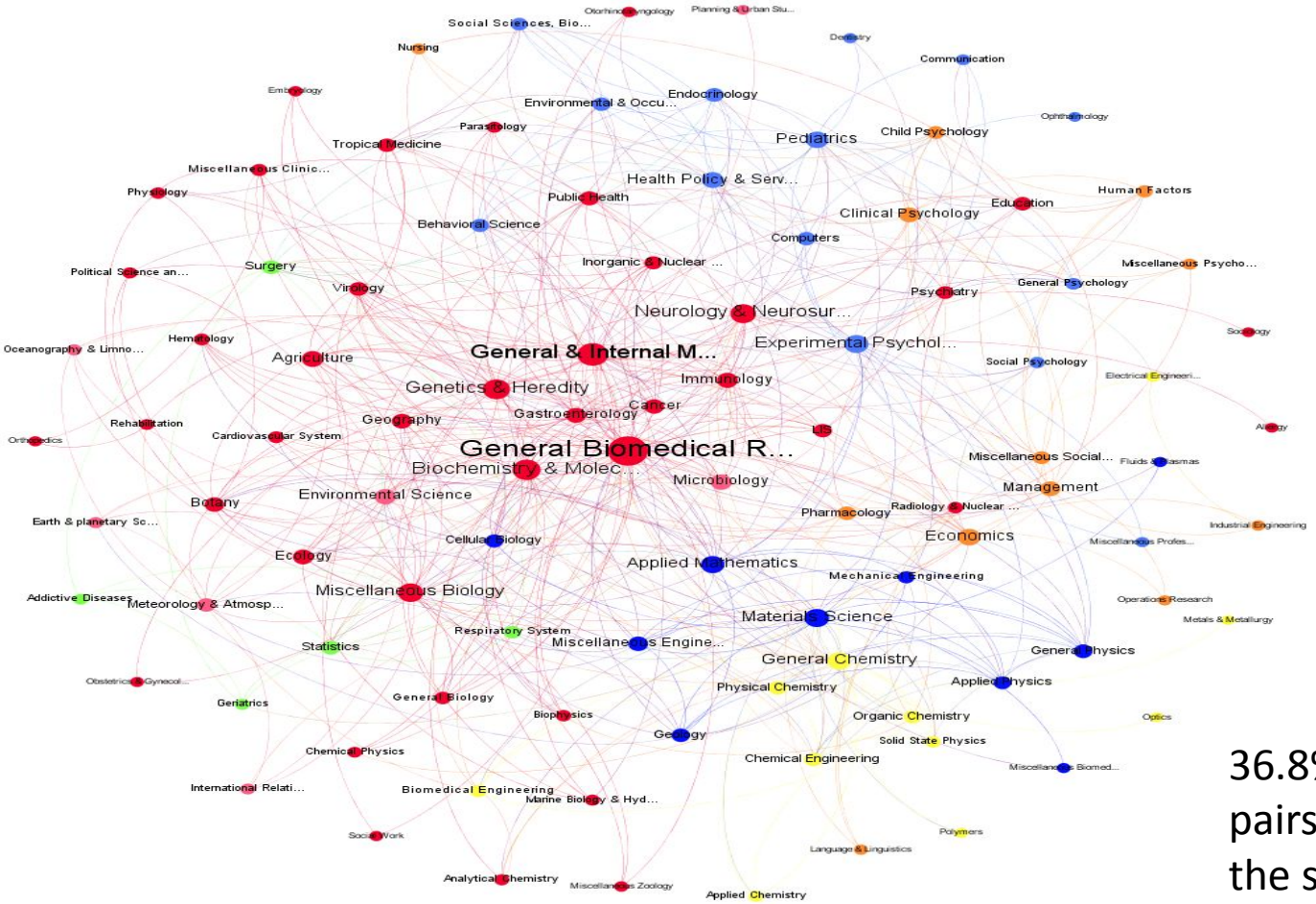
Visibility of OA journals, PLoS ONE, JAMA, …

Co-cited subject categories in WoS



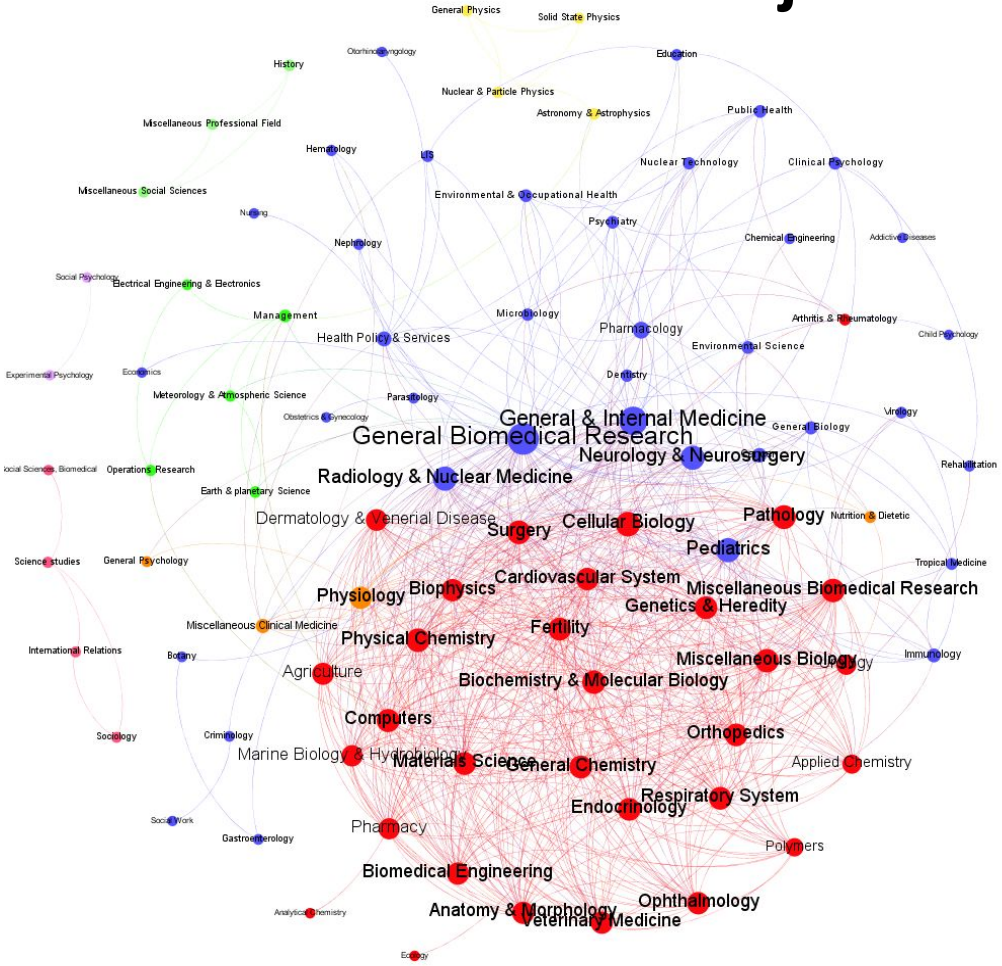
52.4% of co-cited pairs are from the same fields.

Co-read subject categories on Mendeley



36.8% of co-read pairs are from the same fields.

Co-tweeted subject categories



71% of co-tweeted pairs are from the same fields.

Conclusions

- The density and clustering patterns of the three networks differ substantially.
- Citers and Mendeley readers mainly cite/read articles from different journals and subject categories whereas 70% of co-tweeted pairs are from the same journals and subject categories (The likelihood of bot existence, e.g. platform or topic feeds).
- High impact journals are the most active nodes in the citation and readership networks, while open access journals are more visible in the tweet network.
- Comparing co-networks can reveal patterns due to publicity as well as giving insights into the different sources.

3:AM BUCHAREST 2016

Next Week:

More applications