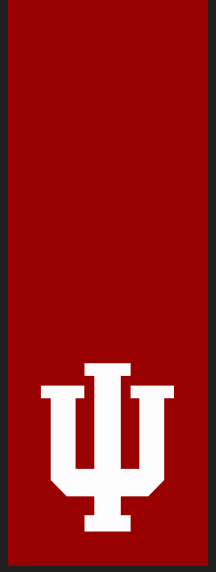


Open Science Forum, 19 October 2016

Correlating Air Transportation with Co-affiliation and Collaboration Data

Xiaoran Yan, Adam Ploszaj, Katy Börner

INDIANA UNIVERSITY BLOOMINGTON

Introduction & methods

Goal & research questions

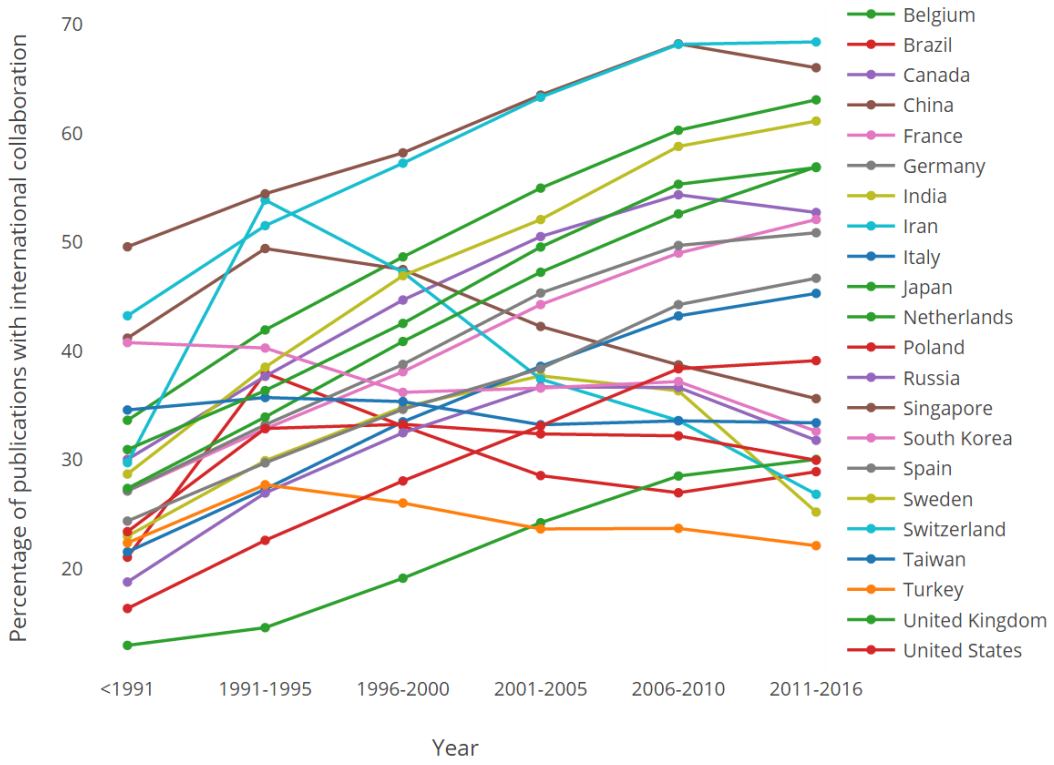
Goal: to capture the interplay of scientific collaboration and transport connectivity on a global scale

Research questions:

1. What are the external scientific collaboration patterns for Indiana University?
2. Are scientific affiliation networks and air traffic networks correlated?
3. Are scientific collaboration networks and air traffic networks correlated?

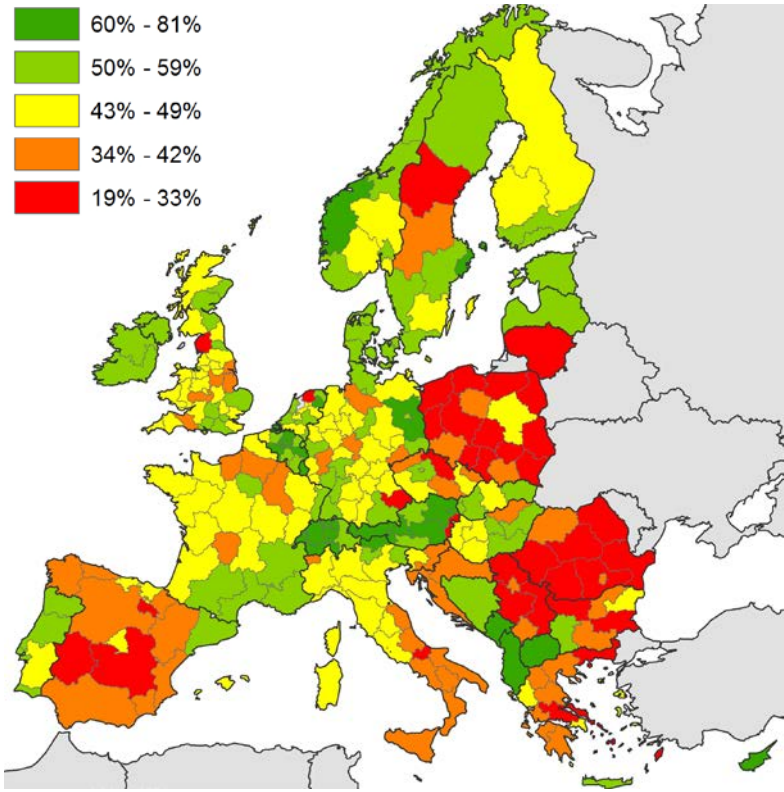
Background: The Fourth Age of Research (Adams 2013)

Publications with international collaboration, data from ResearchGate



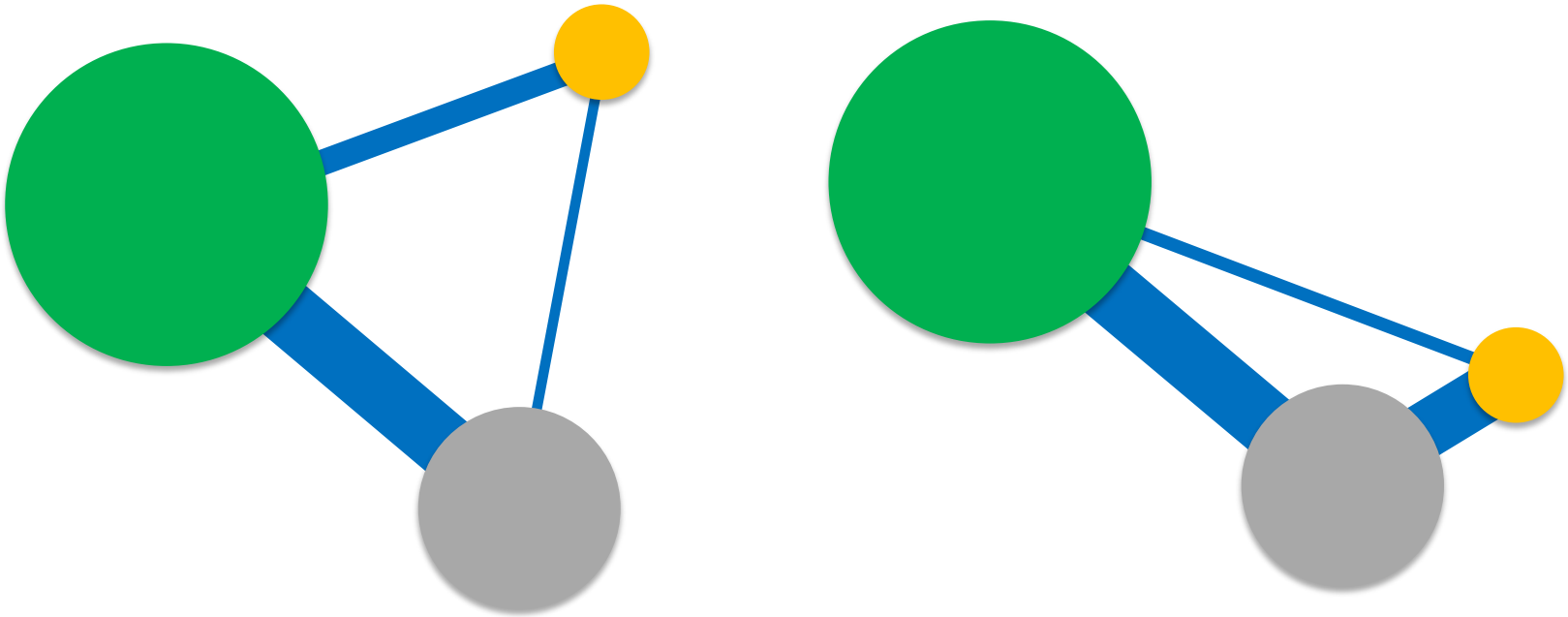
Source: Rikken 2016.

Publications with international collaboration, WoS, 2007-2013



Source: Authors.

Gravity model



- Collaboration as a function of the mass of collaborating entities (e.g. number of publications) and the distance/proximity between them.
- Distance/proximity – not only geographical, but also cognitive, institutional, organizational, social, and economic (Boshma 2005; Fernández et. al 2016).
- Geographical distance and accessibility / connectivity.

Related work

- Research cooperation decreases exponentially with the distance separating the collaborative partners, even when controlling for other factors (e.g. Katz 1994; Fernández et. al 2016).
- Swedish case study: Travel time (road & air) correlated with patents coauthoring (Ejermo, Karlsson 2006).
- Europe: regions/cities with a major international airport are more likely to develop intensive international scientific collaboration (Hoekman et al. 2010).
- US: After Southwest Airlines enters a new route (with lower fares), scientific collaboration increases by 50% (chemistry co-publications, 1991-2012) (Catalini et al. 2016).
- Collaboration vs. co-affiliation (e.g. Sugimoto 2016).

Data sources

IUNI WoS database

- 31,226 IU papers (2008-2013)
- 7,820 papers with co-affiliations
- 27,412 papers with co-authors

Geo coding data

- 2,855 unique cities (ex: Bloomington, IN, USA)

OpenFlights data

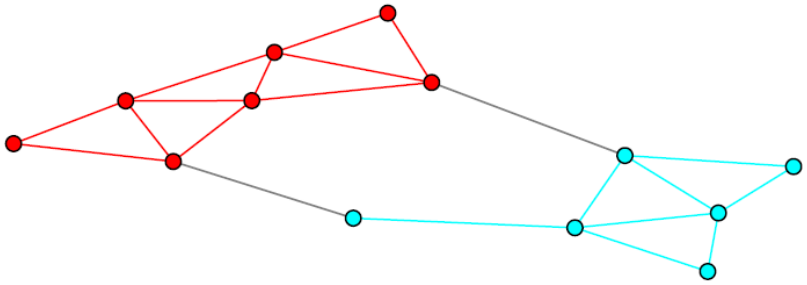
- 3,253 airports and 37,133 weighted flights

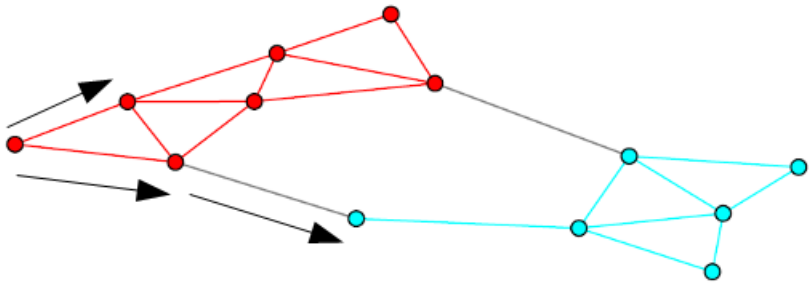
Networks built:

Co-affiliation and Collaboration network for city-level addresses

Air traffic flow network for major airports

Methods
Structure = dynamics + network



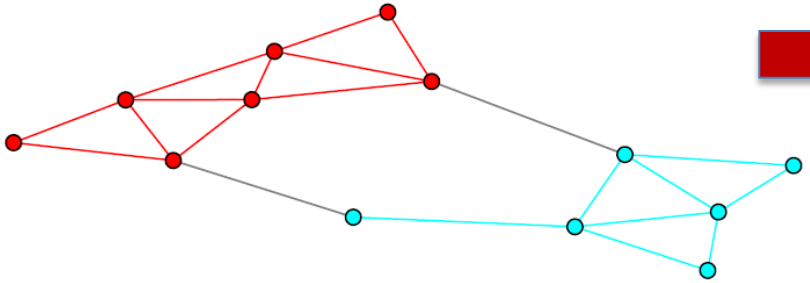


	1	2	...	*n*
1	A_{11}	A_{12}	...	A_{1n}
2	A_{21}	A_{22}	...	A_{2n}
...	...	...	...	...
n	A_{n1}	A_{n2}	...	A_{nn}

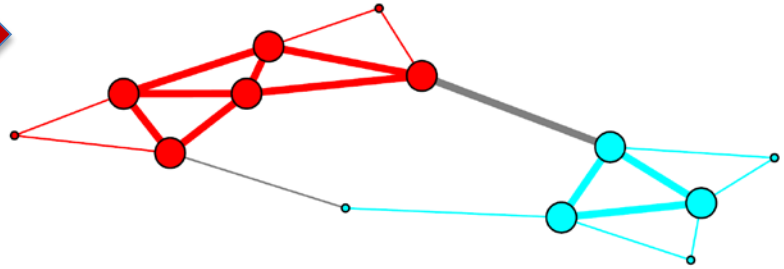
	1	2	...	*n*
1	W_{11}	W_{12}	...	W_{1n}
2	W_{21}	W_{22}	...	W_{2n}
...	...	...	...	...
n	W_{n1}	W_{n2}	...	W_{nn}

Graph transformations

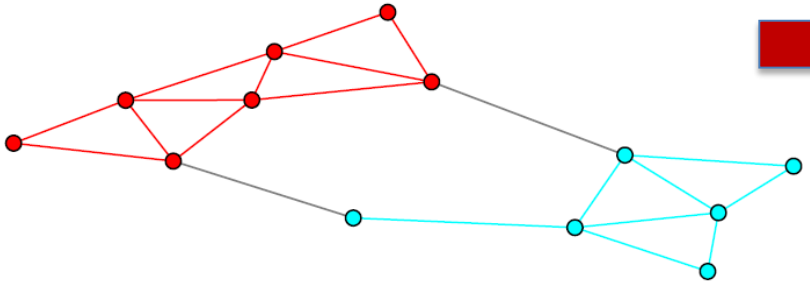
$L=(D_W-BAB)D_W^{-1}$



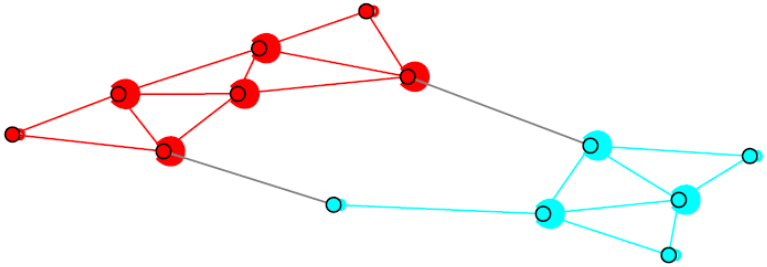
$L=(D-W_1)D^{-1}$



$L=(D-A)D^{-1}T^{-1}$



$L=(D-W_2)D^{-1}$



The Parameterized Laplacian

$$\mathcal{L} = (TD_w)^{-1/2-\rho}(D_w - BWB\ )(TD_w)^{-1/2+\rho}$$

Bias transformation

- Parameterized by B (diagonal)
- W' = BWB or WB for directed graphs

Delay transformation

- Parameterized by T (diagonal)
- Local average delay/rate

Reweighing transformation

- Parameterized by W = R◦A
- Edge specific biases

- $W' = B^b W B^b$
- B for node specific Bias

- $W = c_0 A^a \circ F^f + c_1 1 M^m$
- F, M for edge specific Bias

- Additive bias vs multiplicative bias

The Parameterized Laplacian

The dynamical process

- $B_i = \psi_i$ $$P_{ij} = \frac{A_{ij}}{\lambda}\frac{\psi_j}{\psi_i}.$$
- Maximum entropy RW, over the infinite path distribution
- Stationary $c_i = \psi_i^2$
- Non-conservative

$$\frac{d\boldsymbol{\theta}'}{dt} = (P - \boldsymbol{S})\boldsymbol{T}^{-1}\boldsymbol{\theta}'(t)$$

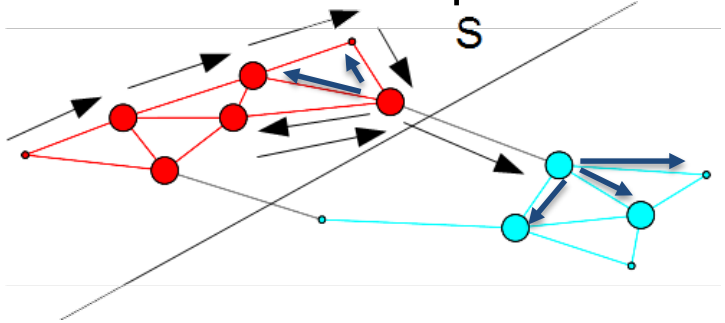
- Other centralities?

Centrality

- $c_i = \psi_i$
- Katz $$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty}\sum_{j=1}^{n}\alpha^k(A^k)_{ji}$$

Community structure

- Bottlenecks in epidemics



Data and results

Data sources

IUNI WoS database

- 31,226 IU papers (2008-2013)
- 7,820 papers with co-affiliations
- 27,412 papers with co-authors

Geo coding data

- 2,855 unique cities (ex: Bloomington, IN, USA)

Air traffic data

- 3,253 airports and 37,133 weighted flights

Networks built:

Co-affiliation and Collaboration network for city-level addresses

Air traffic flow network for major airports

Web of Science dataset

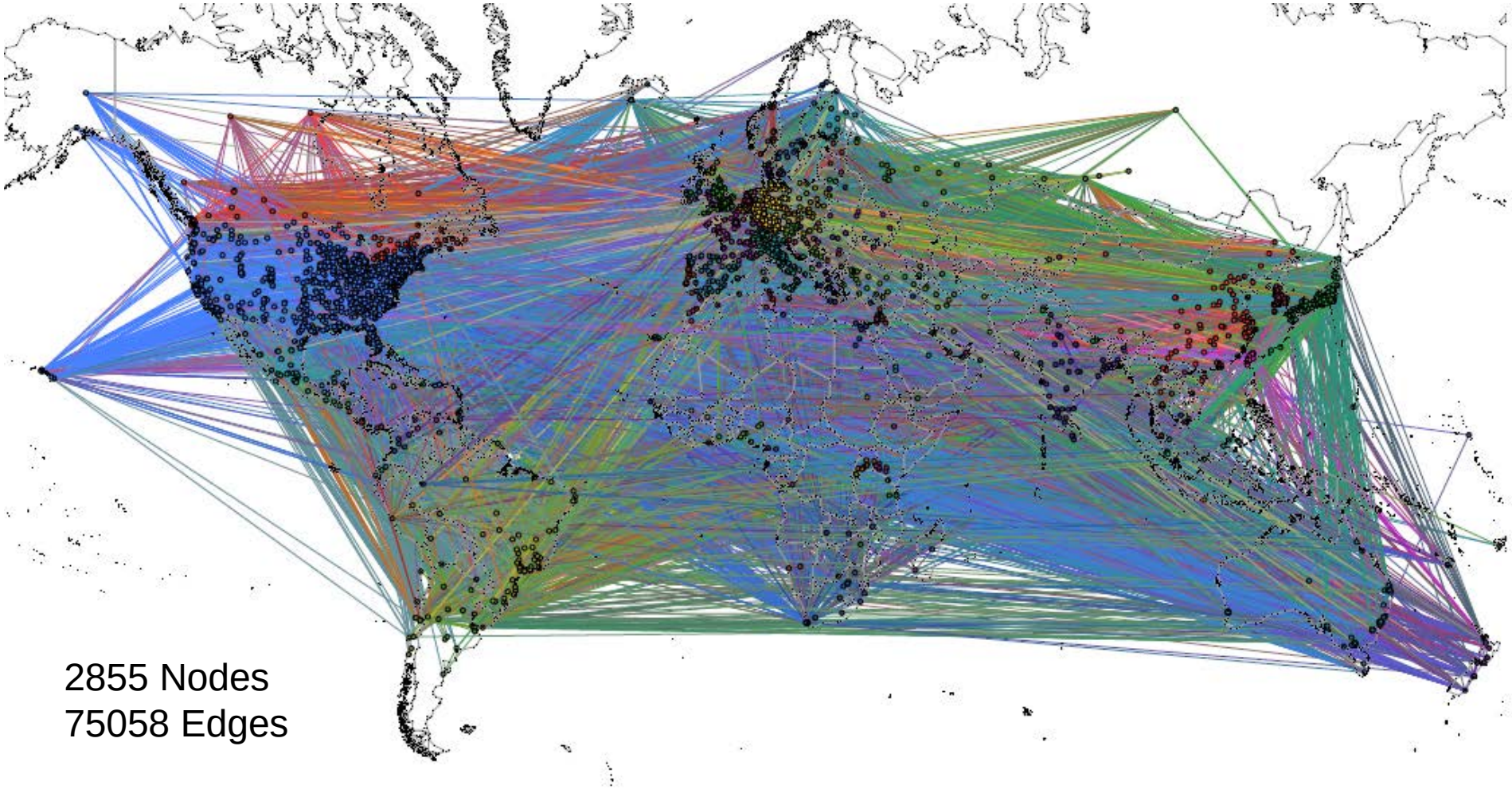
1	id	year	title	journal	fulladdlist	authorlist
2	WOS:0002	1970	Single vibronic	CHEMICAL PHYSI	Indiana Univ, Dept Chem, Bloomington, IN 47401 USA|India	Schuyler, M. W.|Parmen
3	WOS:0002	1970	The O(D-1)+H2(	CHEMICAL PHYSI	Indiana Univ, Dept Chem, Bloomington, IN 47401 USA|India	Hartshorn, Lynn G.|Bair,
4	WOS:0002	1970	Spectroscopic S	INORGANIC CHEI	Univ Queensland, Dept Chem, Brisbane, Qld 4072, Australia	Kitching, William|Doddr
5	WOS:0002	1970	Improved total	CHEMICAL PHYSI	Indiana Univ, Dept Chem, Bloomington, IN 47401 USA;Univ	Bonbam, R. A.|Ng, E. W.
6	WOS:0002	1970	Rate constant f	CHEMICAL PHYSI	Univ Wisconsin, Inst Theoret Chem, Madison, WI 53706 USA	Bernstein, R. B.|Roberts,

Data problems

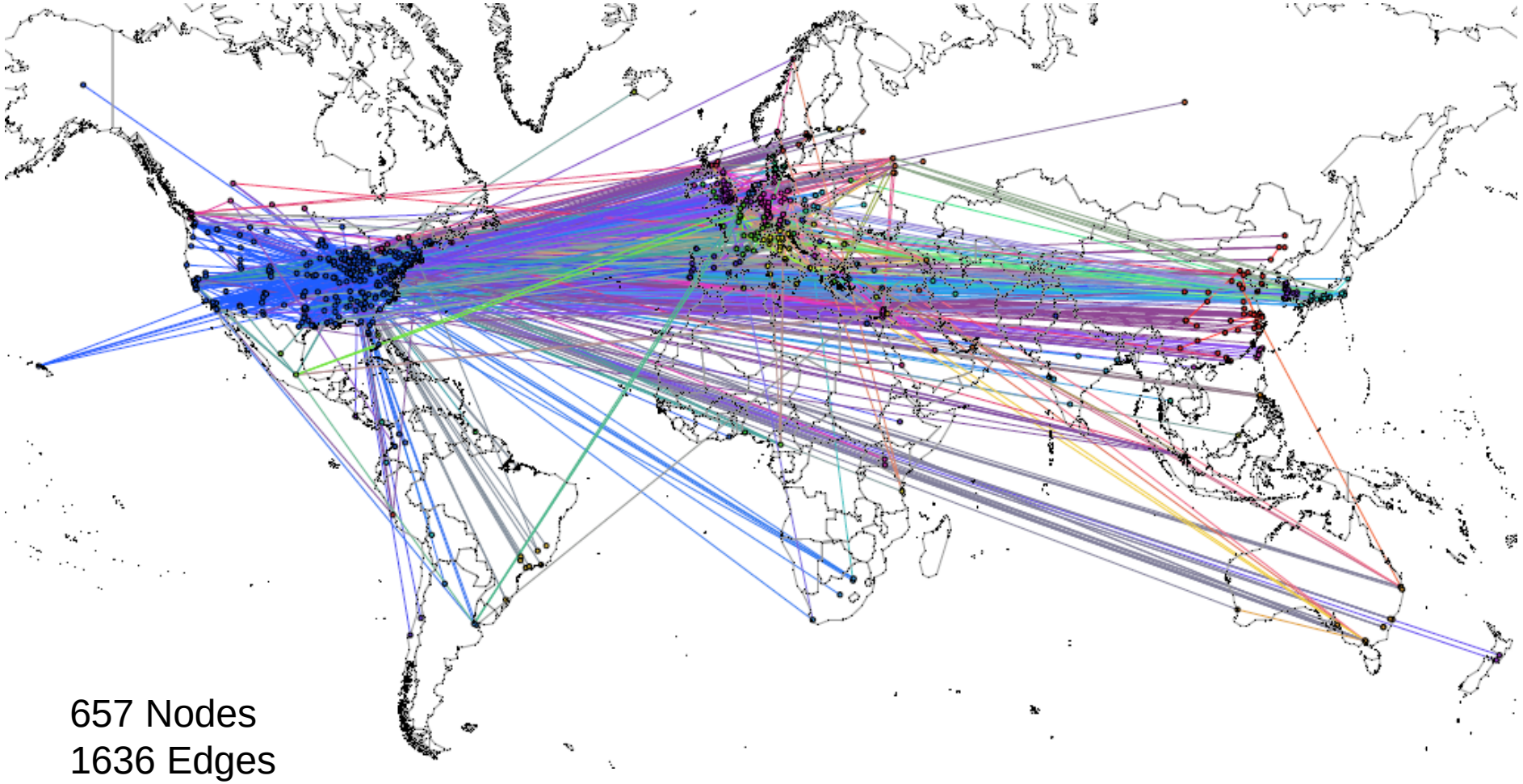
- Lack of data before 2008 with author-address links (1402 total IU papers)
- Noisy address formats, used city-state-country instead
- Author disambiguation, circumvented in this study

1	city	state	country	Latitude_bi	Longitude_bin	County_bing	City_bing	State_bing	Country_bing
2	Millbury	OH	USA	41.5586891	-83.42504883	Wood Co.	Millbury	OH	United States
3	Hamburg	NY	USA	42.7401199	-78.82517242	Erie Co.	Hamburg	NY	United States
4	Harefield	Middx	England	51.5910759	-0.483275145	Harrow	South Harefi	England	United Kingdom
5	Bloomfiel	CT	USA	41.82761	-72.7358017	Hartford Co.	Bloomfield	CT	United States
6	Miki	Kagawa	Japan	36.2816467	139.0772705				Japan

Co-occurrence of city-level addresses for collaborations involving IU authors



Co-occurrence of city-level addresses for IU Authors with Multiple Affiliations



657 Nodes
1636 Edges

Air traffic dataset (OpenFlights)

	id	name	city	country	IATA/FFA	ICAO	Latitude	Longitude	Altitud	Timezone
2	1	Goroka	Goroka	Papua Ne	GKA	AYGA	-6.08169	145.3919	5282	10
3	2	Madang	Madang	Papua Ne	MAG	AYMD	-5.20708	145.7887	20	10
4	3	Mount Ha	Mount Ha	Papua Ne	HGU	AYMH	-5.82679	144.2959	5388	10
5	4	Nadzab	Nadzab	Papua Ne	LAE	AYNZ	-6.56983	146.7262	239	10
6	5	Port More	Port More	Papua Ne	POM	AYPY	-9.44338	147.2201	146	10

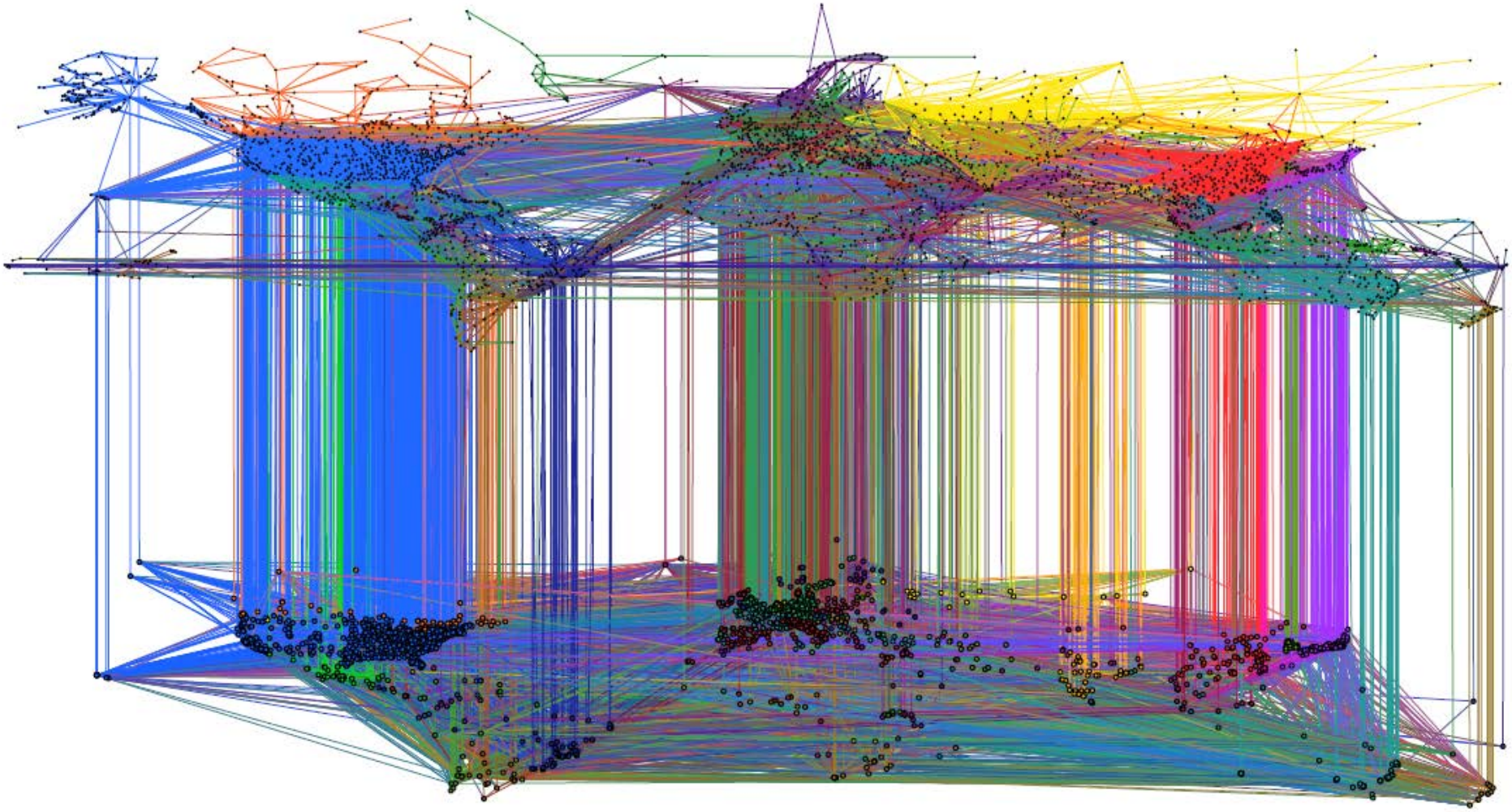
	Airline	AirlineID	Source	SourceID	Destination	Destination	Codeshare	Stops	Equipment
2	PX	328	GKA	1	POM	5		0	DH4 DH8 DH3
3	CG	1308	GKA	1	HGU	3		0	DH8 DHT
4	CG	1308	GKA	1	LAE	4		0	DH8
5	CG	1308	GKA	1	MAG	2		0	DH8
6	CG	1308	GKA	1	POM	5		0	DH8

	Equipment	Manufacturer	Type/Model	Wake	Seats
2	EM2	EMBRAER	EMB 120 Brasilia	L	40
3	DH8	De Havilland Canada	DHC-8 Dash 8	M	120
4	320	Airbus	A320-100/200	M	150
5	321	Airbus	A321-100/200	M	200
6	744	Boeing	747-400	H	416

Air traffic data network

3253 Nodes
37133 Edges

Bimodal network of 2863 unique city-level affiliations with closet airports

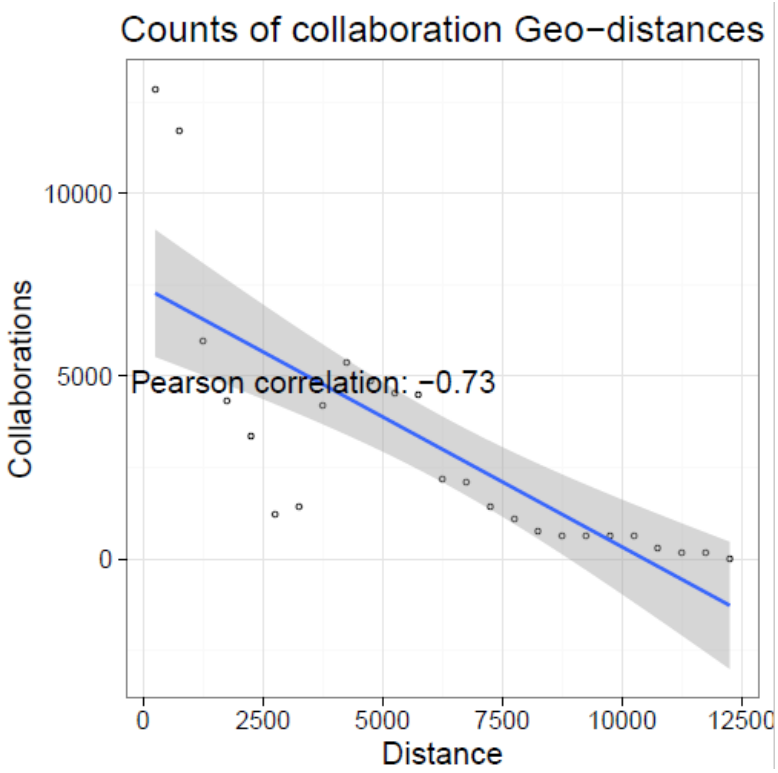


City level collaboration pairs

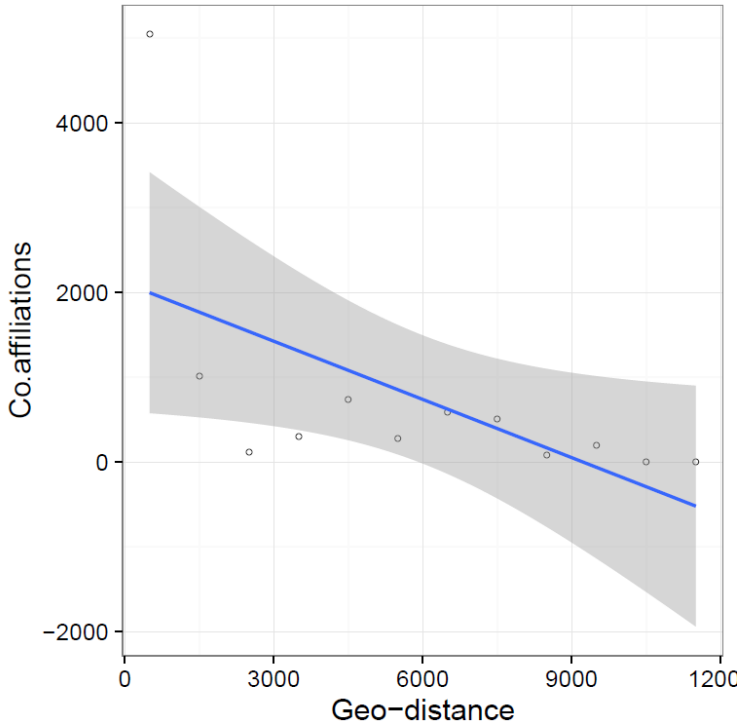
1	Source	Target	Id	Dist2AS	Dist2A	FlightSeats	Co-affiliatio	GeoDist	Collaborations
2	n0	n1	e0	7.875831	14.12155	1.39E+08	0	951.0674	148
3	n0	n2	e1	42.90588	110.8321	6.05E+08	7	480.9978	281
4	n0	n3	e3599	11.16308	23.19697	6.05E+08	49	438.2402	689
5	n0	n4	e16835	12.56916	27.07884	9.05E+08	0	344.6902	126
6	n0	n5	e3611	13.95766	30.91222	4.1E+09	0	2261.481	61
7	n0	n6	e7879	11.32607	23.64692	1.19E+09	0	335.1899	96
8	n0	n7	e16984	9.47158	18.52708	3.69E+08	0	435.382	15
9	n0	n8	e7923	4.800921	5.632372	1.57E+09	0	2161.833	11
10	n0	n9	e13027	15.18406	34.29804	7.89E+08	0	658.4138	67
11	n0	n10	e1186	12.40009	26.61208	4.02E+08	10	1299.628	169

2855 Nodes
75058 Edges

Pair of attributes: Geo distance Vs collaboration/co-affiliation

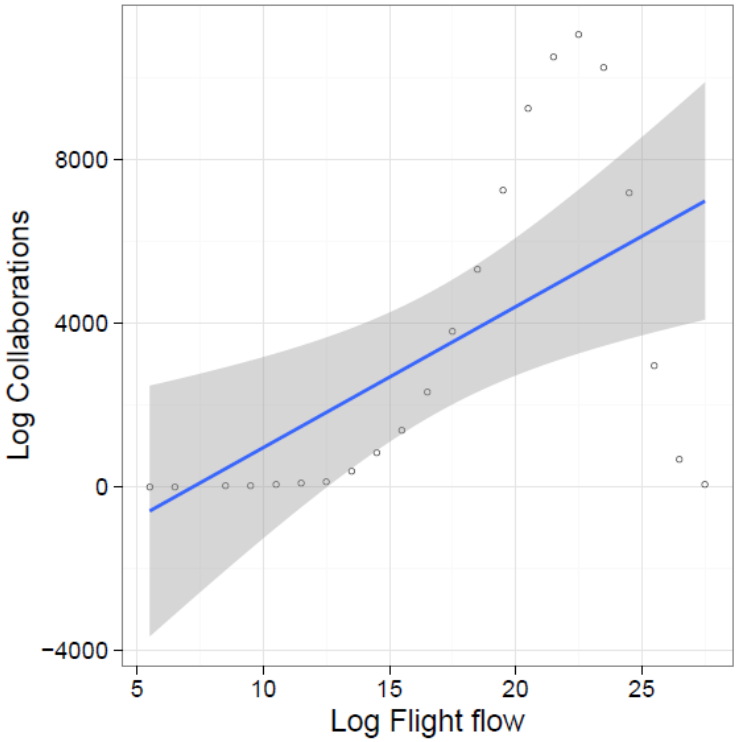


Pearson's coefficient: -0.73

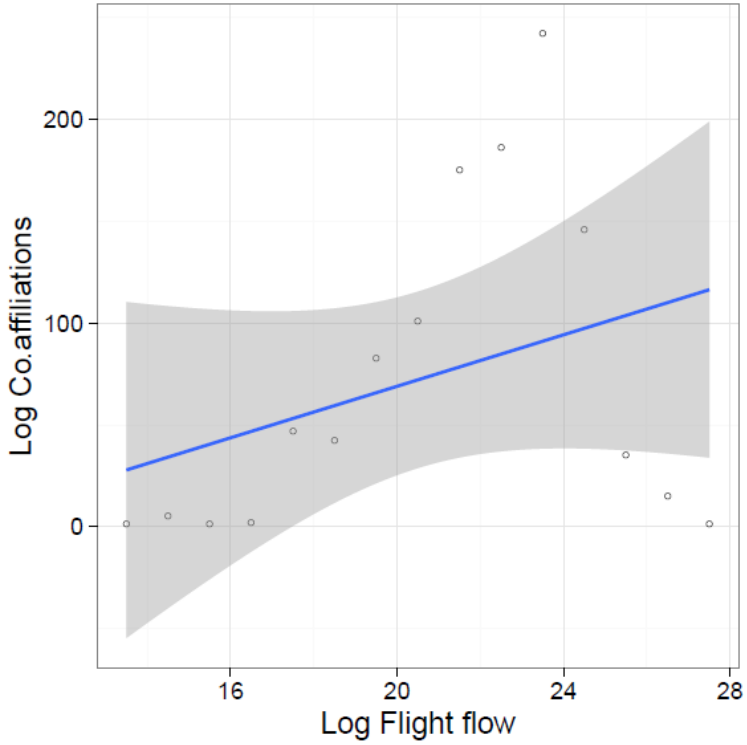


Pearson's coefficient: -0.59

Pair of attributes: Air traffic flow Vs collaboration/co-affiliation

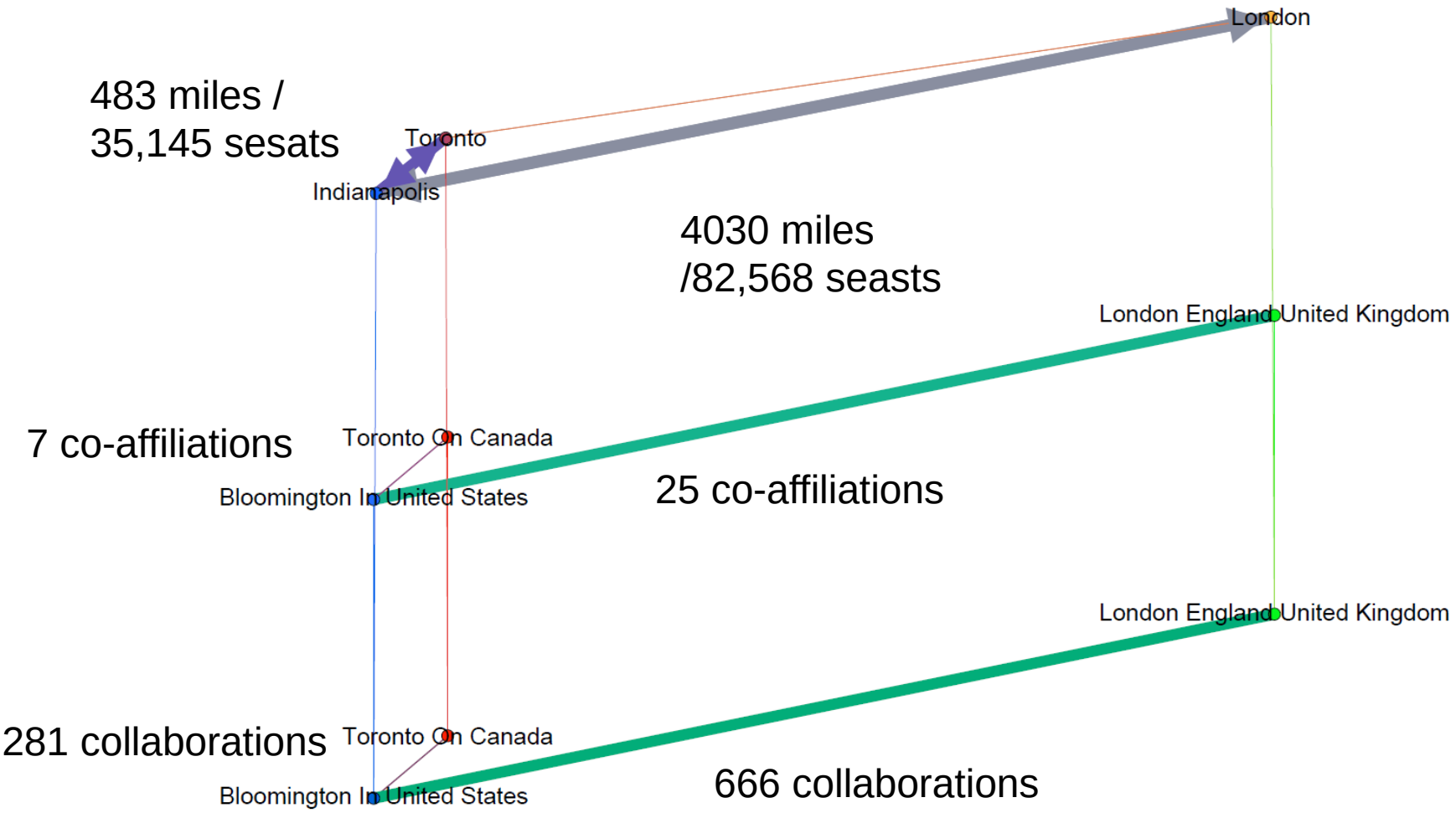


Pearson's coefficient: 0.59



Pearson's coefficient: 0.35

$$W = G \circ F^{-1} \circ M^{-1}$$

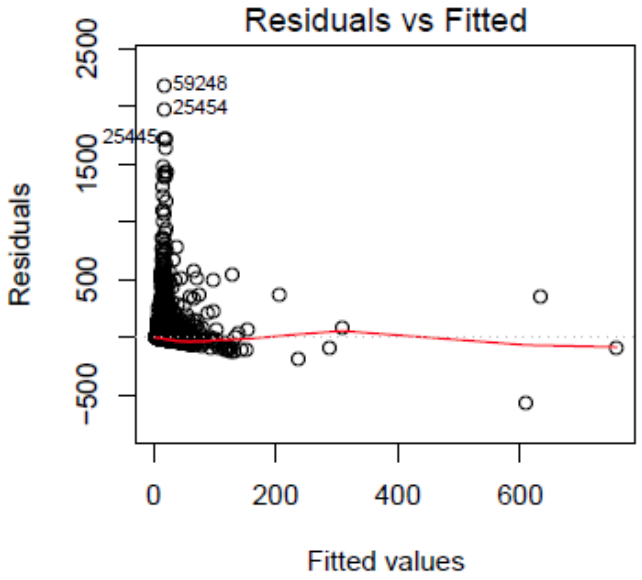


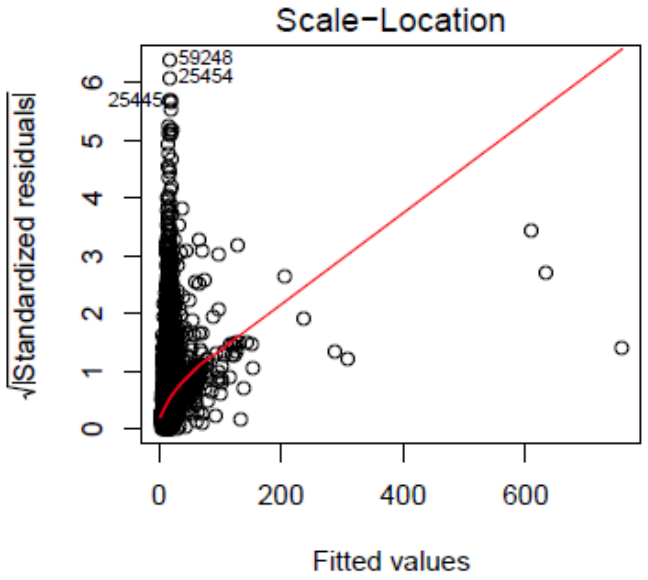
Regression analysis

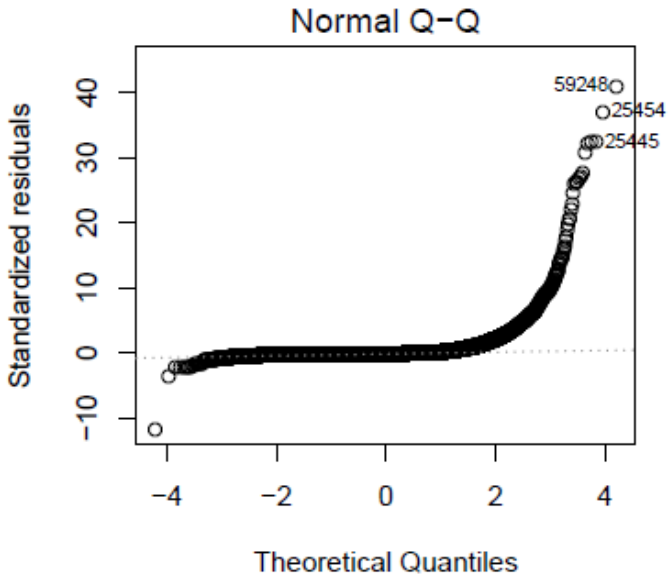
$$W' = B^b W B^b,\ W = G^g \circ F^f \circ M^m$$

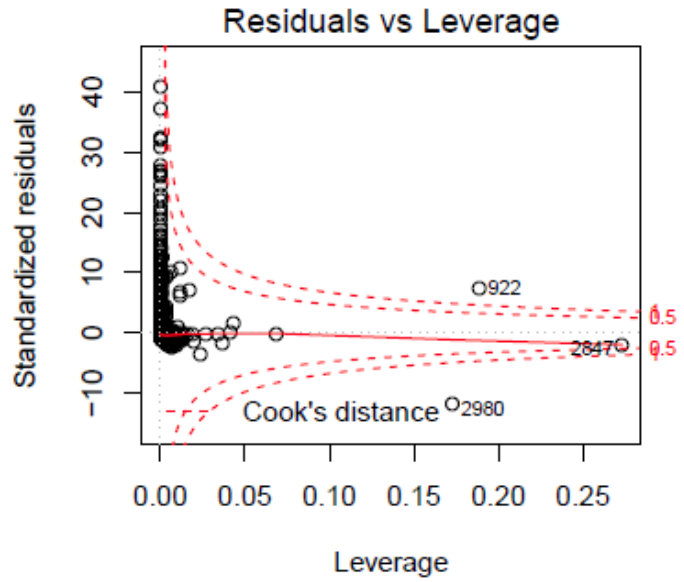
$$log(W'_{uv}) = b*log(B_u B_v) + a*log(A_{uv}) + f*log(F_{uv}) + m*log(M_{uv})$$

```
Residuals:
    Min      1Q  Median      3Q     Max
-576.35  -14.50  -11.39   -4.65 2174.18

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.272e+01  7.736e-01  29.365   <2e-16 ***
GeoDist         -1.463e-03  1.350e-04 -10.839   <2e-16 ***
Dist2A          -5.772e-04  6.756e-04  -0.854    0.393
FlightSeats      6.026e-11  7.099e-12   8.489   <2e-16 ***
Co.affiliations  4.499e+00  1.681e-01  26.770   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.36 on 40241 degrees of freedom
Multiple R-squared:  0.0222,    Adjusted R-squared:  0.0221
F-statistic: 228.4 on 4 and 40241 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
Fitted values
59248
25454
25445
2500
1500
500
−500
0
200
400
600


Scale−Location
√|Standardized residuals|
Fitted values
59248
25454
25445
6
5
4
3
2
1
0
0
200
400
600


Normal Q−Q
Standardized residuals
Theoretical Quantiles
59248
25454
25445
40
30
20
10
0
−10
−4
−2
0
2
4


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
922
2847
2980
0.5
0.5
40
30
20
10
0
−10
0.00
0.05
0.10
0.15
0.20
0.25

Conclusions & further steps

Conclusions & further steps

- Affiliation network correlates with air traffic network stronger than collaboration network. (Possible explanation: co-affiliation needs more fiscal presence than collaboration.)
- Air traffic network geodistances and collaboration patterns.

Further steps

- Comparative case studies:
  - (1) IUB + U Mich, Ann Arbor + Cornell U, Ithaca,
  - (2) Organizations from Europe and/or China.
- Adding explanatory variables.
- Adding more detailed air traffic data (for the US available from U.S. Department of Transportation).

Web of Science as a Research Dataset

Date:

November 14-15, 2016

Meeting Place:

Social Science Research Commons (SSRC),
Woodburn Hall, Room 200
1100 East Seventh Street
Bloomington, IN 47405

Web Indiana University Campus Map »

Workshop Goals

This practical workshop brings together data scientists and data stewards from research centers that are using the Web of Science™ at scale. We will explore WoS from the perspective of a research dataset and work together on practical ways to better support our research in the future. While the main focus will be on the Web of Science, the results should be extensible to all similar metadata aggregations. This unique focus—bringing data stewards and data scientists from these centers together to work on shared needs in tandem with the Web of Science team—will enable us to redefine and fully repurpose WoS to fit our research goals. We intend to launch an ongoing community in which we will learn techniques and develop tools to improve the data that underlies our research.

Advance Preparations

- Data stewards will provide a short profile of how WoS as a dataset is being implemented in the context of their research center/university and the technical, content, and other challenges they are facing.
- Researcher data scientists will prepare a short profile of current research projects leveraging the WoS dataset, focusing on key challenges such as linking, disambiguating, mining, etc. that, if solved, would offer greater research opportunities.

Organizers:

Katy Börner

Victor H. Yngve Distinguished Professor of Information Science, Department of Information and Library Science, School of Informatics and Computing, Indiana University, Bloomington; Director, Cyberinfrastructure for Network Science Center & Curator of Mapping Science exhibit, Bloomington, IN
katy@indiana.edu

Eamon Duede

Executive Director, Knowledge Lab. Administrator, Metaknowledge Research Network, University of Chicago
eduede@uchicago.edu

James Pringle

Head of Industry Development & Innovation at Thomson Reuters IP & Science