Who is the Erdős of Economics?

Thomas Krichel[1,2,3]

[1]Long Island University [2]Novosibirsk State University [3]Open Library Society

Guildford 2012–03–16

thanks to

- Rob Witt and the Department
- Open Society Institute and JISC
- the RePEc gang, especially Christian Zimmermann
- Vincent J. Berton, Jr

also thanks to

- Society for Economic Dynamics
- Dept. of Economics at WUStL.
- William Waites
- Francisco Javier Moraiz Gargallo

Erdős

- Paul Erdős (1913–1996) was a very prolific mathematician.
- He wrote 1500+ papers with 511 co-authors.

Erdős number

- Published mathematicians keep an Erdős number.
- It the number of co-authors separating them from Erdős.
- There is some prestige in a low number.

issues

- Beyond the anecdotal I am not aware of actual evidence that Erdős was ever the most central mathematician.
- We would find it difficult to relate this to economics.

what do we need

- data, with detour on current work PART 1
- methods PART 2
- calculation infrastructure PART 3

data

- Document data. We get this from RePEc.
- Author identification data. We get that from the RePEc Author Service RAS.
- The latter is crucial.

without RAS data

- There are many ways one can express an author's name. Just look at my Novosibirsk visitor page.
- There are many authors who share a name expression, e.g. Michael Devereux.
- Automated methods are not successful.

RePEc

- RePEc is misunderstood as a repository.
- in fact it is a collection of 1300+ institutional, discipline-specific repositories.
- some of the largest come from publishers.
- roughly 1.1 million bib items

personal registration

- I created the RePEc Author Service RAS in 1999.
- Initially called “HoPEc”.
- The first programmer was Markus J.R. Klink.

stage i

- At initial registration, the author gives personal information.
  - email address
  - name and name variations

name variation

- It is assumed that a paper may have been written by an author if it has matched one of her name variations.
- RAS also performs some fuzzy searches offline to spot spelling mistakes.

stage ii: claiming papers

- This is the heart of RAS.
- Authors claim or disclaim papers that carry a name variation of their.
- There is an email alert service for new matching papers.

success of RAS

- over 30k authors registered
- from an old independent list of top 1000 economists over 80% are registered.

a RePEc for all disciplines

- RePEc bibliographic data → 3lib
- RePEc Author Service → AuthorClaim
- EDIRC → ARIW

3lib

- 3lib is an initial attempt at building an aggregate of *freely* available bibliographic data.
- It's a project by OLS sponsored by OKFN.
- 40+ million records from the usual: PubMed, OpenLibrary, DBLP, RePEc and institutional repositories

3lib elements

- The data elements in 3lib are very simple
  - title
  - author name expressions
  - link to item page on provider site
  - identifier
- 3lib is meant to serve AuthorClaim.

AuthorClaim

- AuthorClaim is an authorship claiming service for 3lib data.
- It lives at http://authorclaim.org.
- It uses the same software as the RePEc Author Service, called ACIS.
- It is running since early 2008.

advantages of AuthorClaim

- Bulk data freely available
- Robust and simple design.
- Avoids authors to work with many claiming systems.

AuthorClaim data

- ftp://ftp.authorclaim.org
- CC0
- more than 100 profiles, growing slowly.

so far so good

- I should not to talk about AuthorClaim but about a services that we can build when we have identified authors.
- When we have this data, we can find out who has been writing papers with whom.
- In other words we can study the co-authorship network.

PART 2: methods

- I will illustrate the methods by looking at a very small example set of data.
- This is gathered from AuthorClaim and E-LIS

E-LIS

- It is a subject-based eprints archive for library and Informations Science.
- Founded by Antonella de Robbio in 2003.
- Over 10k papers

E-LIS plus AuthorClaim

- Some folks registered with AuthorClaim and claimed E-LIS papers
- This forms a very small real-world dataset used to illustrate the methods.
- Data shown here were correct as of 1 November 2011.

445 papers claimed by 36 authors (1)

- 72 Tomas Baiget
- 61 Ulrich Herb
- 43 Antonella De Robbio
- 39 Thomas Krichel
- 26 Andrea Marchitelli & fernanda peset,
- 20 Ross MacIntyre

445 papers claimed by 36 authors (2)

- 16 Dirk Lewandowski
- 15 Bożena Bednarek-Michalska
- 14 Lidia Derfert-Wolf
- 11 Zeno Tajoli & Imma Subirats
- 9 Derek Law & Emma McCulloch & Philipp Mayr
- 8 Jeffrey Beall

445 papers claimed by 36 authors (3)

- 7 Nuria Lloret Romero
- 6 Benjamin John Keele
- 5 Adrian Pohl & Maria Francisca Abad-Garcia
- 4 Walther Umstaetter
- 3 Andrea Scharnhorst & Jose Manuel Barrueco & Thomas Hapke &

445 papers claimed by 36 authors (4)

- 3 Christian Hauschke & Klaus Graf
- 2 Frank Havemann & Eberhard R. Hilf & Bhojaraju Gunjal & Chris L. Awre
- 1 Loet Leydesdorff & Peter Bolles Hirtle & Alexei Botchkarev & Christina K. Pikas & Oliver Flimm & Sridhar Gutam

co-authorship

- When two registered author claim to have authored the same paper, we say that they are co-authors.
- The authorship relationship creates a link between the two authors.
- The link is symmetric.

58 papers have been co-claimed by 16 co-authors

- 12 fernanda peset
- 10 Tomas Baiget
- 8 Imma Subirats
- 6 Antonella De Robbio
- 4 Nuria Lloret Romero

58 papers have been co-claimed by 16 co-authors

- 2 Andrea Marchitelli & Ulrich Herb & Ross MacIntyre & Boena Bednarek-Michalska & Thomas Krichel & Dirk Lewandowski & Lidia Derfert-Wolf
- 1 Derek Law & Emma McCulloch & Sridhar Gutam & Philipp Mayr

network and components

- When we start with one co-author, and we move to her co-authors, what other authors can be reach?
- We call the authors we can reach by starting from any one of them by following co-authorship relationships a component of the network.

network components (1)

- Scottish: Derek Law & Emma McCulloch
- Polish: Boena Bednarek-Michalska & Lidia Derfert-Wolf
- German: Dirk Lewandowski & Sridhar Gutam & Philipp Mayr

network components (2)

- Giant: Andrea Marchitelli & Ulrich Herb & Thomas Krichel & Antonella De Robbio & fernanda peset & Imma Subirats & Ross MacIntyre &Nuria Lloret Romero & Tomas Baiget

the giant component

- The giant component is larger than all other components together.
- Most real existing networks have a giant component.
- As the network grows, older small components join the giant component and new small components are created.

centrality

- Who is at the most central author in E-LIS?
- The answer is that it depends on how we measure centrality.
- Two measures are commonly used
  - closeness centrality
  - betweenness centrality
- Both depend on a measure of distance

distance

- To understand that we need a measure of distance.
- We say that two authors have distance one if they are co-authors.
- We say that two authors have distance two if they are not co-authors, but have a common co-author.
- etc

shortest paths

- In order to find the distance between two authors, we have to evaluate all possible paths between them.
- We need to find shortest paths between. There are well-known algorithms to find them.
- The distance is the length of the shortest path.

distances for Imma Subirat

- Tomas Baiget 1
- Antonella De Robbio 1
- Ulrich Herb 2
- Thomas Krichel 1
- Nuria Lloret Romero 2
- Andrea Marchitelli 2
- Ross MacIntyre 2
- fernanda peset 1

distances for Ulrich Herb

- Tomas Baiget 1
- Antonella De Robbio 3
- Thomas Krichel 2
- Nuria Lloret Romero 3
- Andrea Marchitelli 4
- Ross MacIntyre 4
- fernanda peset 2
- Imma Subirats 2

closeness centrality

- The average distance of Imma is much smaller than the average distance of Ulrich.
- In fact, we can calculated to average distance of the every author from all other authors.
- This is what we call closeness centrality of an author.

diameter

- This is the length of the longest shortest paths between any two authors.
- In our network the diameter is four.
- This much smaller than the number of authors in the giant component (9).
- We say that our network has the small world property.

shortest paths from Tomas Baiget (1)

- $\longrightarrow$ Thomas Krichel
- $\longrightarrow$ fernanda peset $\longrightarrow$ Nuria Lloret Romero
- $\longrightarrow$ fernanda peset
- $\longrightarrow$ Imma Subirats $\longrightarrow$ Antonella De Robbio $\longrightarrow$ Ross MacIntyre

shortest paths from Tomas Baiget (2)

- ⟶ Ulrich Herb
- ⟶ Imma Subirats ⟶ Antonella De Robbio
- ⟶ Imma Subirats ⟶ Antonella De Robbio ⟶ Andrea Marchitelli
- ⟶ Imma Subirats

shortest paths from Antonella De Robbio (2)

- $\longrightarrow$ Imma Subirats $\longrightarrow$ fernanda peset $\longrightarrow$ Nuria Lloret Romero
- $\longrightarrow$ Imma Subirats
- $\longrightarrow$ Imma Subirats $\longrightarrow$ Tomas Baiget $\longrightarrow$ Ulrich Herb
- $\longrightarrow$ Imma Subirats $\longrightarrow$ Tomas Baiget

shortest paths from Antonella De Robbio (2)

- ⟶ Imma Subirats ⟶ fernanda peset
- ⟶ Andrea Marchitelli
- ⟶ Ross MacIntyre
- ⟶ Thomas Krichel

s. p. from Ross MacIntyre

- ⟶ Antonella De Robbio ⟶ Imma Subirats ⟶ fernanda peset ⟶Nuria Lloret Romero
- ⟶ Antonella De Robbio ⟶ Imma Subirats ⟶ fernanda peset
- ⟶ Antonella De Robbio ⟶ Imma Subirats
- ⟶ Andrea Marchitelli

s. p. from Ross MacIntyre

- $\longrightarrow$ Antonella De Robbio $\longrightarrow$ Imma Subirats $\longrightarrow$ Tomas Baiget $\longrightarrow$ Ulrich Herb
- $\longrightarrow$ Antonella De Robbio $\longrightarrow$ Thomas Krichel
- $\longrightarrow$ Antonella De Robbio $\longrightarrow$ Imma Subirats $\longrightarrow$ Tomas Baiget
- $\longrightarrow$ Antonella De Robbio

what do the paths tell us?

- Some authors are appear more often as intermediaries than others.
- We can evaluate the number of times an author appears as an intermediary.
- This is the betweenness centrality of an author.
- A number of authors have zero betweenness. They are called marginal authors.

summary

- We build a network.
- We find two ways to evaluate authors
  - closeness
  - betweenness
- Now let us look at the results.

ranking for closeness (1)

rank	name	closeness
1	Imma Subirats	1.5
2	Antonella De Robbio	1.75
2	Tomas Baiget	1.75
2	Thomas Krichel	1.75

ranking for closeness (2)

rank	name	closeness
5	fernanda peset	1.875
6	Andrea Marchitelli	2.5
6	Ross MacIntyre	2.5
8	Ulrich Herb	2.625
9	Nuria Lloret Romero	2.75

ranking for betweenness

rank	name	betweenness
1	Antonella De Robbio	2.7
1	Imma Subirats	2.7
3	Tomas Baiget	2.025
4	fernanda peset	1.575

Andrea Marchitelli, Ross MacIntyre, Nuria Lloret Romero, me and Ulrich Herb are marginal.

the missing bit

- In this simple model, we have only looked at one shortest path.
- In fact there can be multiple shortest paths.
- I did not show them up there. I think they are not there is that sample scenario.

path multiplicity

- They appear as soon as we have a rectangular collaboration.
  - A write with B and D.
  - B write with A and C.
  - C write with D and B.

what paths?

- It is possible to generalize the binary model to incorporate collaboration strengths.
- The most sensible approach that I have seen comes from a paper by Mark Newman.

weighted model (of Mark Newman)

- Let a paper be written by A authors.
- If two authors have collaborated in the paper, their collaboration strength is augmented by $1/(A-1)$.
- The distance between two authors is one divided by the collaboration strength.

desirable properties

- The sum of the collaboration strength of an anther is the sum of the papers an author has collaborated on.
- for more see M. E. J Newman, Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, Physical Review E. Vol. 64, 016132, 2001.

undesirable properties

- It violates triangular inequality.
- I first saw this when I looked up my connection to Joe Pearlman on a full set of calculations using a network build on this weighted collaboration.

Pearlman problem example (1)

- P_1: A_1 A_2
- P_2: A_1 A_3
- P_3: A_2 A_3 A_4 A_5 A_6

Pearlman problem example (2)

- $s(A_1, A_2) = 1$
- $s(A_1, A_3) = 1$
- $s(A_2, A_3) = 1/4$

Pearlman problem example (3)

- $d(A_1, A_2) = 1$
- $d(A_1, A_3) = 1$
- $d(A_2, A_3) = 4$

the problem

- Shortest path from A_2 to A_3 is: $A_2 \longrightarrow A_1 \longrightarrow A_3$.
- This is all right for an academic paper.
- For a running services, it is not acceptable.

to proceed

- I calculate all shortest binary paths.
- I evaluate the weighted length of all.
- I eliminate all shortest paths that don't have minimum weighted length.

PART 3: calculation infrastructure

- There are 20k+, meaning that there are a lot of paths.
- I have a special machine fricka.openlib.org that does only do the calculations. It is behind a firewall.
- A RePEc machine compiles the information about RAS every day.

updating

- A set of processes takes turn, takes the oldest updated author.
- We calculate the shortest “paths” file for starting from an author.
- We calculate an “inter” file with intermediate authors used by the author.
- We calculate a “report” file.

time

- At this time, over 1 million seconds between current time and average file.
- The average retrieved path is 500k seconds old.

retrieval

- If we need to show the shortest path between two author, we have two sources.
- We use the most recent source.

generic implementation

- ICANIS is as generic software that is (with one minor fudge) adaptable to many networks.
- It has complete separation between results and presentation.
- It is used to calculate the AuthorClaim, E-LIS and RAS networks.

RAS network

- http://collec.repec.org
- This is not the machine doing the computations.

to do

- full code documentation
- home page search
- produce summary statistics of the network

the end

- Thank you for your attention
- http://openlib.org/home/krichel