Statistical Analysis of Networks

Lecture 8 – Data Collection

NETWORK DATA COLLECTION

Network graph or adjacency matrix are constructed from measurements taken on a set of entities and their interactions (a system of interest):

- very often choices have to be made on
  - Entities and scale (individuals, organizations, devises, users, topics, words, papers, journals,…)
  - Interactions (often not necesserely or uniquely defined) (structural variables in NA)
- choices (definitions) influence network construction and network results
  - General consideration: the possible choices depend on the specific discipline/context of interest
    - technological, social, biological, information, knowledge, … networks are usually very different requiring specific compentences on the problem(s) of interest and the kind of availble data
- different empirical methods used to measure network interactions
  - two of the most important:
    - direct questioning of subjects
    - use of archival records (online databases)

Different Designs

- The broad majority of network studies use either “whole-network’’ (complete network) or “egocentric” designs.

    - Whole network design examine sets of interrelated objects or actors that are regarded for analytical purposes as bounded social collectives (one-mode networks)
    - Ego-centric is similar to standard statistical analysis, where the survey units (ego) provide answers on their relations with alters and those of the alters (eventually)

- Data sets with two sets of objects – termed *two-mode* – are becoming common

NETWORK BOUNDARY

units to be considered as the set of nodes

- characteristics and size
  - specific (sometimes small) study: an organization, department, student class, Zachary's karate club,…

- characteristics with respect to the interaction
  - the specified set provides all the information on the system of relations?

- population (or partial) and sampled network data
  - *enumerated* (population) data: information on what is considered a population **as a whole** (with time and/or geographical or other constraints)
  - *sampled* data: the system of interactions is build from the **sampling of units** from a population
    - specific design of network sampling: *snowball sampling*, a technique developed for studying hidden populations (e.g. drug users) that relies on interactions for its operation, starting to collect information on some nodes and then looking to their neighors and so on in an iterative process. Variants are used in other context too (www and html pages)
- data quality issues: missing data and measurement errors

DIRECT QUESTIONING (MAINLY IN SNA)

Direct questioning of subjects (nodes or informants)

(face to face or using CAI methods of data collection)

Name generator approach:

- information on network partners (other nodes in the network) are obtained directly from the actors
- an item (or series of items) invite respondents to name others with whom they have contact of a *specified kind (usually directed relation are produced)*

NAME GENERATOR

- A survey item (or series of items) that invite respondents to name others with whom they have contact of a specified kind
- Approach can be used in both whole and ego-centric network design

Two main possible choices:

- Obtaining (relational) information on a pre-specified list of actors (or alters in the ego-centric-design)

- Obtaining names (or categories of alters) without any pre-specified suggestion

Usually name generator is based on free recalling of the respondents without any limit of the alters that can be mentioned

McCarty (2002) proposed a free recall of respondents with a fixed number of choices

NAME GENERATOR

- A survey item (or series of items) that invite respondents to name others with whom they have contact of a specified kind
- Approach can be used in both ego-centric network design

Example: friendship networks among schoolchildren, children are asked to complete a questionnaire that included items worded as follows:

My best friend at 'School Name' is:
My second-best friend at 'School Name' is:
My third-best friend at 'School Name' is:
...
My eighth-best friend at Junior High School is:

fixed choice list: **eight-best** friend (limited out-degree but not in-degree)

free choice**:** no limit
Ideally all students within the school would be surveyed
Note that the survey specifically asks children to name only friends within the school (no external friends: choices about it)

NAME GENERATOR

Often respondents are asked not just to name those with whom they have ties but to describe the nature of those ties as well

Example: study on a group of medical doctors, respondents were asked the following questions:

Who among your colleagues do you turn to most often for advice?

With whom do you most often discuss your cases in the course of an ordinary week?

Who are the friends among your colleagues who you see most often socially?

- fixed choice: maximum three doctors's names could be given in response to each question
- questions on *several types* of interactions: data on *several different networks*
- information on respondent attributes are usually collected as well
  - age, gender, income, education (assortative mixing by attributes)

(see also Lazega - *lawyers dataset:*
coworker network, advice network, friendship network)

NAME GENERATOR

Main disadvantages of network studies based on direct questioning:

- high demanding
  - CAI methods of data collection can help
  - usually limited to few hundreds of units

- typical biases affecting statistical surveys
  - accurate and careful use of survey methods

ARCHIVAL DATA (widespread availability of online databases)

Florentine families (Padgett, 1994)

investigation of historical records to determine which among the families had

- marriage ties
- trade relations (a clear definition is needed)
- other forms of social contact with one another (a clear definition is needed)

Harry Potter books

- character network in *the Goblet of Fire* (4th book)

https://anthonybonato.com/2016/08/03/social-networks-in-novels-and-films/

- support network in *the Philosopher's Stone* (and the other books)

https://www.stats.ox.ac.uk/~snijders/siena/HarryPotterData.html

https://towardsdatascience.com/explore-harry-potter-via-a-dynamic-social-network-of-characters-f5bed9a39f01: 5 different applications: characters' importance, narrative structural change, community detection, summarisation and book comparison

Game of throne

- HBO TV show character network in season 1
  - TV series based on the novel series entitled *A Song of Fire and Ice* by George R.R. Martin.
  - co-appearance: if two characters share a scene in the TV show, they are recorded as sharing a connection.

https://gameofnodes.wordpress.com/2015/05/06/game-of-nodes-a-social-network-analysis-of-game-of-thrones/

ARCHIVAL DATA - HARRY POTTER AND THE GOBLET OF FIRE (4TH BOOK)

BONATO *ET AL.* (2016)

Network statistics

	# Nodes	Avg. Degree	Avg. Weighted Degree	Diameter	Edge Density	Avg. Distance	Clust. Coeff.
Goblet	62	18.55	305.29	2	0.304	1.69	0.746

edges are *co-occurrence*: characters appearing together in the text.

edge weight: # of co-occurrence between two names.

No direct reading of the book: authors used an extraction algorithm searching for character names being **fifteen or fewer words apart in the text**.

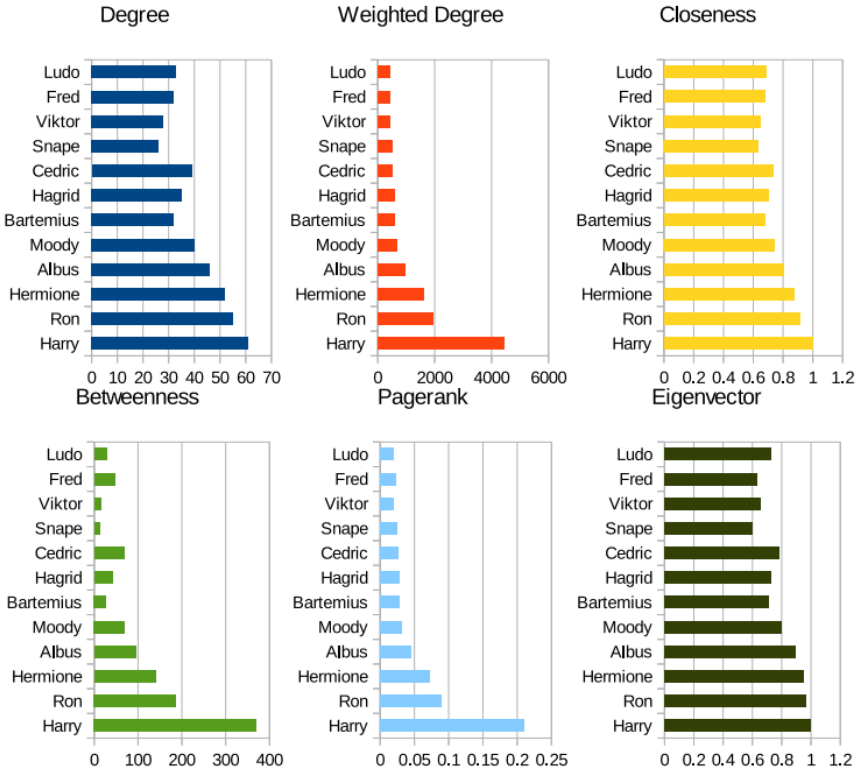


Fig. 3. Centrality measures for *Harry Potter and the Goblet of Fire*.

ARCHIVAL DATA - HARRY POTTER AND THE GOBLET OF FIRE (4TH BOOK)

BONATO *ET AL.* (2016)

Colors represent **communities** (picked out by **Louvain algorithm**)*:* Hogwarts, the Dursleys, the Weasleys, Slytherin, and the friends Seamus and Dean.

Louvain algorithm: each node is first assigned to a community on its own. Then, each node is moved to the community in which it achieves the highest contribution to the ***modularity*** (fraction of the edges in the network that connect nodes within-community minus its expected value in the case of a network with edges placed at random).

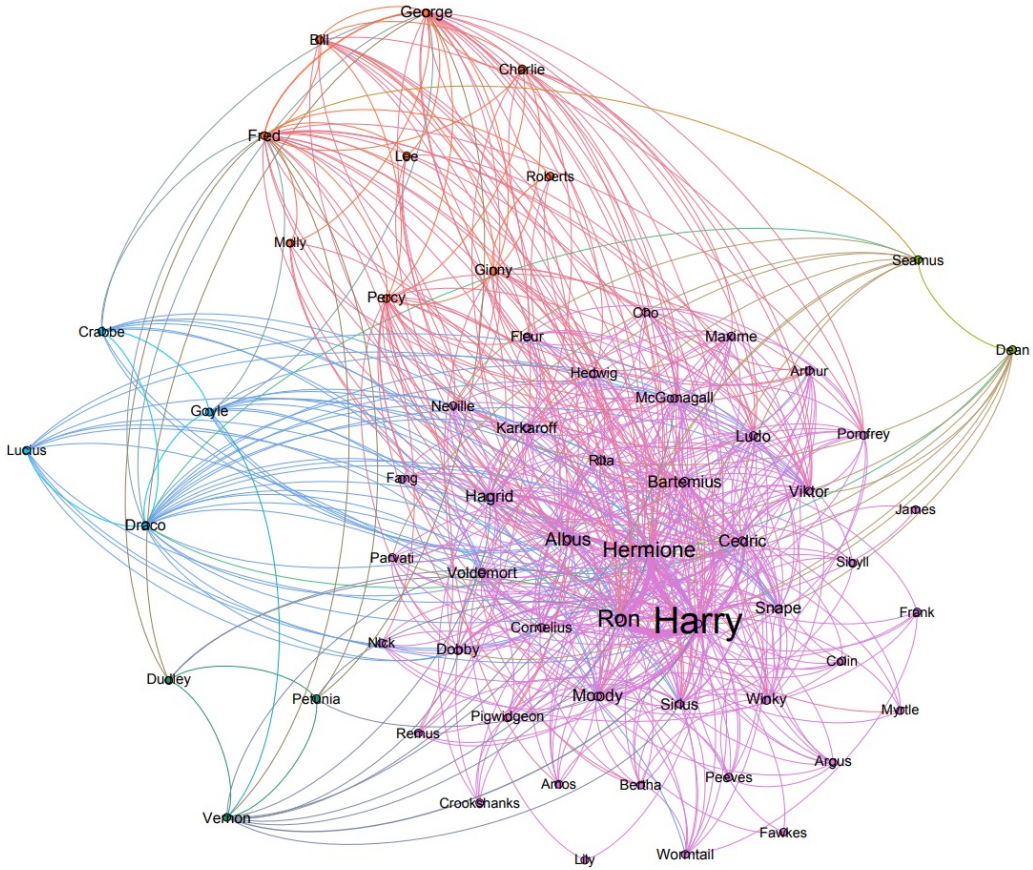


Fig. 2. The character network for *Harry Potter and the Goblet of Fire*. Each community is represented by a distinct color. The thickness of an edge is scaled to its weight, and the size of a name is scaled to the Pagerank score.

ARCHIVAL DATA - HARRY POTTER AND THE Philosopher's Stone (1ST BOOK)

Bossaert and Meidert (2013)

Edge: Peer support

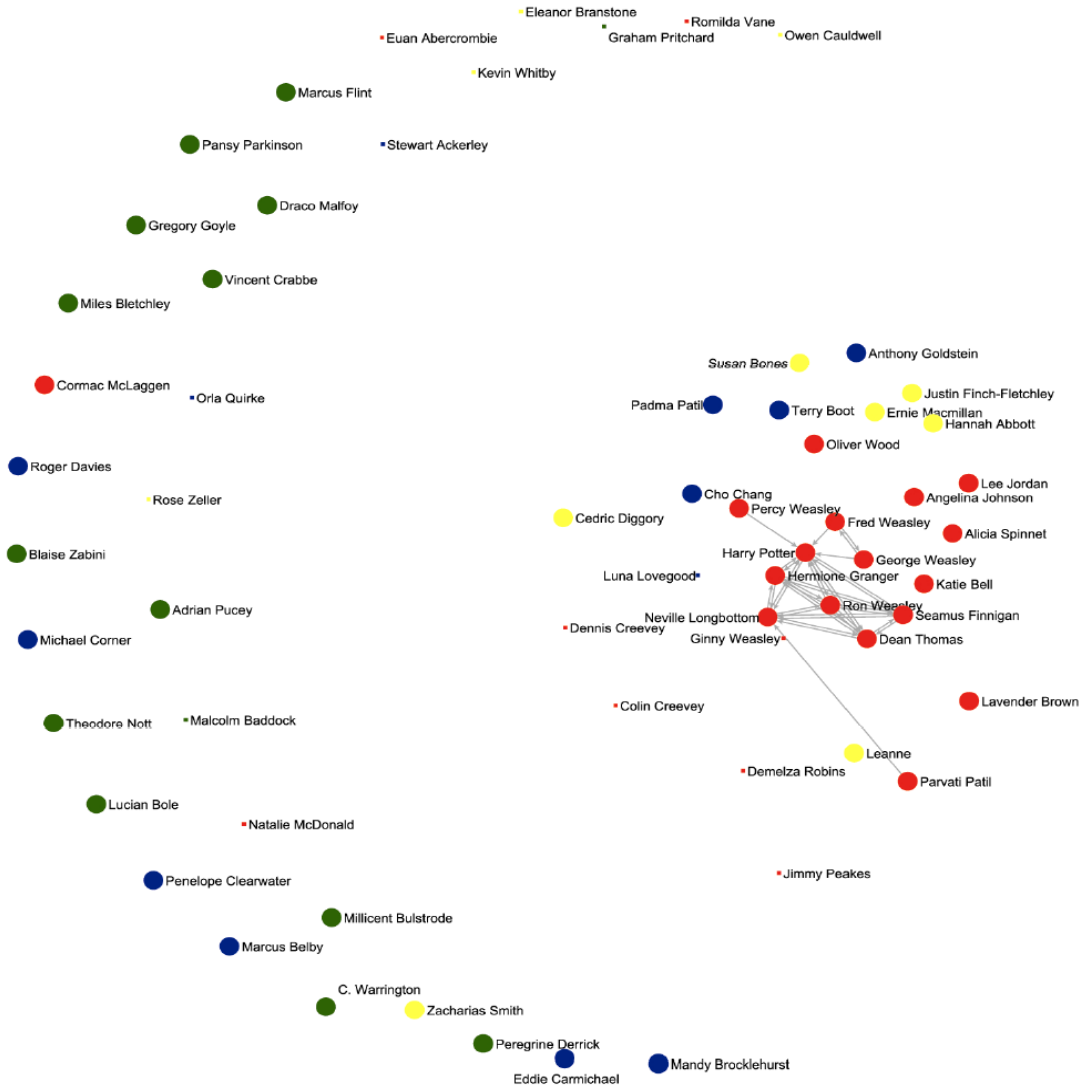
Contact between the 64 Hogwarts students was coded as peer support:

1) Student A supports student B emotionally
2) Student A gives students B instrumental help
3) Student A gives student B information to help student B
4) Student A praises student B

two extra conditions:

a) contact between students was only coded if the peer support was offered voluntarily

b) only interactions occurring between two living characters, attending Hogwarts at the same moment, were coded as peer support.



ARCHIVAL DATA - GAME OF THRONE

https://networkofthrones.wordpress.com/

From Book to Network https://networkofthrones.wordpress.com/from-book-to-network/

Link two characters each time their names (or nicknames) appear **within 15 words of one another** (a network is generated for each book)

ARCHIVAL DATA - GAME OF THRONE

From Book to Network https://networkofthrones.wordpress.com/from-book-to-network/

Interaction could be direct or undirect

Types of interactions:
- Two characters appearing together in the same location
- Two characters in conversation
- One character talking about another character
- One character listening to a third character talk about a second character
- A third character talking about two other characters (undirected)
- And so on…

automatic removal of edges of weight 1 or 2 (considered incidental)

Different thresholds for interaction (10 words, 15 words, 20 words):
tuning to 15 words produced the most reasonable network for George R. R. Martin's novels.

'*Of course, determining when two characters have interacted is easier said than done. The more effort you put into identifying interactions, the better your network*'

Archival data - Game of throne

From Book to Network https://networkofthrones.wordpress.com/from-book-to-network/

Three main sources:

1. **The Books.** Our most naive effort was to parsed the books, looking for character names by keeping track of capitalized words. This was a messy process, but did yield a reasonable start.
2. **Web Scraping.** Using data science techniques, we processed A Wiki of Ice and Fire, a fan-created site for the books and the TV series. This was our best source of truth, since character articles contain lists of aliases and titles (come capitalized, some not). This site also includes a list of books that reference the character, but we found this information was not very reliable for the minor characters
3. **First Appearance Spreadsheet.** Leo King's A Song of Ice and Fire Character Spreadsheet. Leo compiled a list of first appearances of characters. This was helpful for us, but not quite as useful as the fan-created wiki.

Characters

Archival data - Game of throne

https://networkofthrones.wordpress.com/

Disambiguation: to solve ambiguous references.

- Many characters share the same names (Jon, Walder, Brandon) and titles (king, queen, maester).
- Does a given appearance of the name “Jon” refer to Jon Snow or Jon Arryn?
- When someone references “the king,” then this reference could resolve to any number of people (Aegon, Robert, Joffrey, Robb, Stannis…), depending on who is speaking and the context in which they are speaking. Likewise, “dwarf” usually refers to Tyrion, but there are instances where it does not.

Disambiguation was a labor-intensive **manual** process (also automatic techniques)

ARCHIVAL DATA - GAME OF THRONE

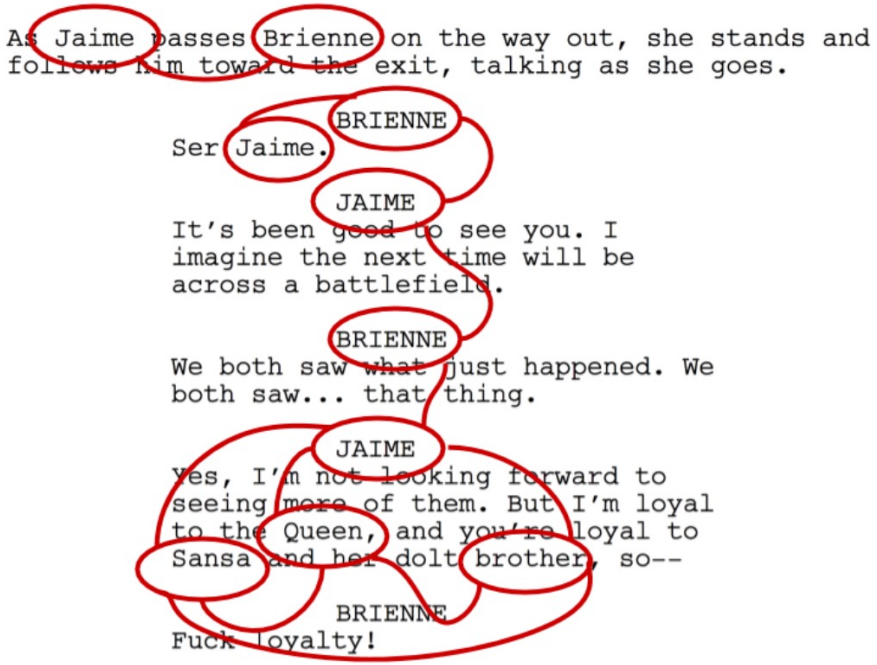
https://networkofthrones.wordpress.com/

From Script to Network

The scripts for the show are not publicly available (except for those that have been submitted for Emmy consideration).

Fan-authored scripts available on genius.com.
These scripts start with the closed caption subtitles, and are then organized into scenes and embellished with stage directions.
Some seasons are incomplete or missing, so some scripts have to be created ad hoc.

Scripts processing: adding a link between Character A and Character B when:
1.They appear in a scene together
2.They appear in a stage direction together
3.They exchange dialog
4.Character A mentions Character B
5.Another character mentions Character A and Character B together.



ARCHIVAL DATA - AFFILIATION NETWORKS

An important special case of the reconstruction of networks from archival records is the affiliation network:

- nodes are connected by co-membership of groups of some kind
  - the Southern Women Study (Davis et al. 1941): data from newspaper reports of social events and the "groups" were the sets of individuals (women) who attended particular events.
  - CEOs of companies in Chicago in the 1970s and their social interaction via clubs that they attended: CEOs are the actors and the clubs are the groups (Galaskiewicz, 1985)
  - Co-authorsip networks: some extremely large affiliation networks
    - an actor is an academic author and a group is the set of authors of a paper.
    - well documented networks in the last few years with the appearance of extensive online bibliographic resources covering many scientific disciplines

Example: co-authorship network of Italian academic statisticians

(not very large but with some interesting aspects to deal with)

Archival data – co-authorship network

Substantial body of works on the analysis of co-authorship patterns at national and international level:

- Physics and Biomedical research (Newman, 2004; Barabasi, 2002) ⇒ MEDLINE and Spires
- Economics (Goyal, 2006; Maggioni and Uberti, 2011) ⇒ Econlit
- Sociology (Moody, 2004; Ferligoj and Kronegger, 2011, 2012) ⇒ Sociological Abstracts and national archive (COBISS database)

Common aims:

- understanding networks properties through SNA
- implications of collaboration patterns on the evolution over time of topics and methods
- evaluation of scientific productivity and quality

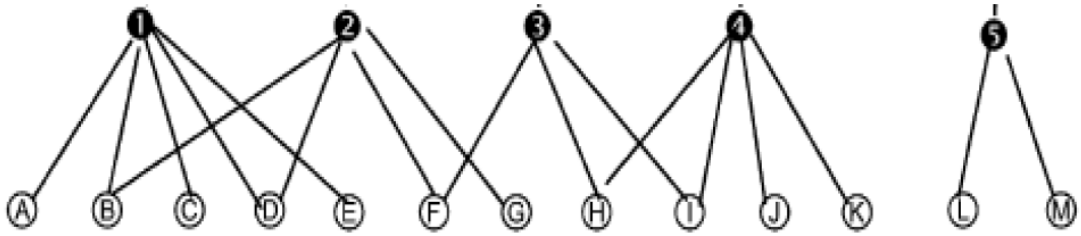
Network topology approach

Emergence of peculiar structures (i.e. theoretical structure with well-defined topological and relational properties) governing collaboration behavior by detecting network statistical main properties

Archival data – co-authorship network

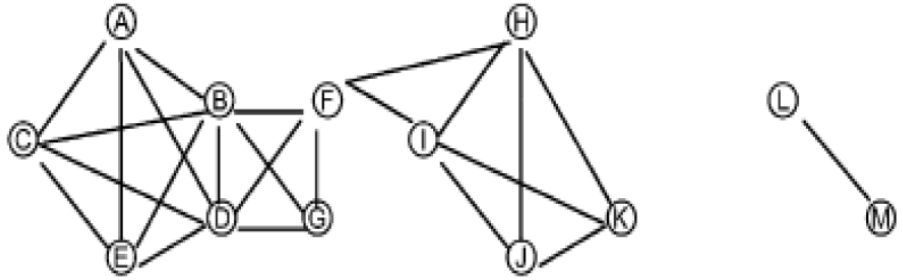
Two-mode network

$\mathbf{AP}_{(n \times m)}$ two sets of nodes:
n **A**uthors $\times$ m **P**apers



One-mode undirected network

$\mathbf{AA}_{(n \times n)} = \mathbf{AP} \times \mathbf{AP}^T$
authors as nodes, papers as links
$a_{ij} \geq 1$ if i-th author have written at least a paper with j-th author



Usually the adjacency matrix **AA** is analysed as a binary network (0/1)

CO-AUTHORSHIP NETWORK OF ITALIAN STATISTICIANS

Co-authorship patterns in Statistics, focusing on:

academic statisticians in Italy (792 grouped in 5 subfields, at March 2010, Today 836)

1. to investigate if the emerging pattern of diffusion of statistical knowledge among Italian scholars resembles the typical one observed in the literature in Natural or in Social Sciences

Collaboration on a specific discipline

Seminal studies on co-authorship patterns: based on international databases containing mainly high-impact publications about the discipline

Collaboration on a specific scientific community (target population)

- query (international) bibliographic archives (selected publications)
- individual scientic CVs
- local/national bibliographic archives

good coverage of whole research products of each scientist

CO-AUTHORSHIP NETWORK OF ITALIAN STATISTICIANS

Why (Italian) Statistics?

1. co-authorship behaviour in Statistics has not yet been investigated

2. Statistics presents some characteristics common to Natural sciences as well as Social sciences, playing a central role in all sciences in view of the importance of statistical methods in everyday applications

3. no unique archive for publications in Statistics: interest to trace this specific target population in distinct data sources

New project (funded by MUR): University of Trieste, University of Milano Bicocca, University of Catania

- Academic Italian Statisticians, Sociologists and scholars in Marketing and Management
- Main datasource for network construction: Scopus