Class 2 -
Basics of Networks - collecting, modeling and representing

Course: Computational Network Analysis

Prof. Dr. Claudia Müller-Birn

Institute of Computer Science, «Human-Centered Computing»

Feb 23, 2016

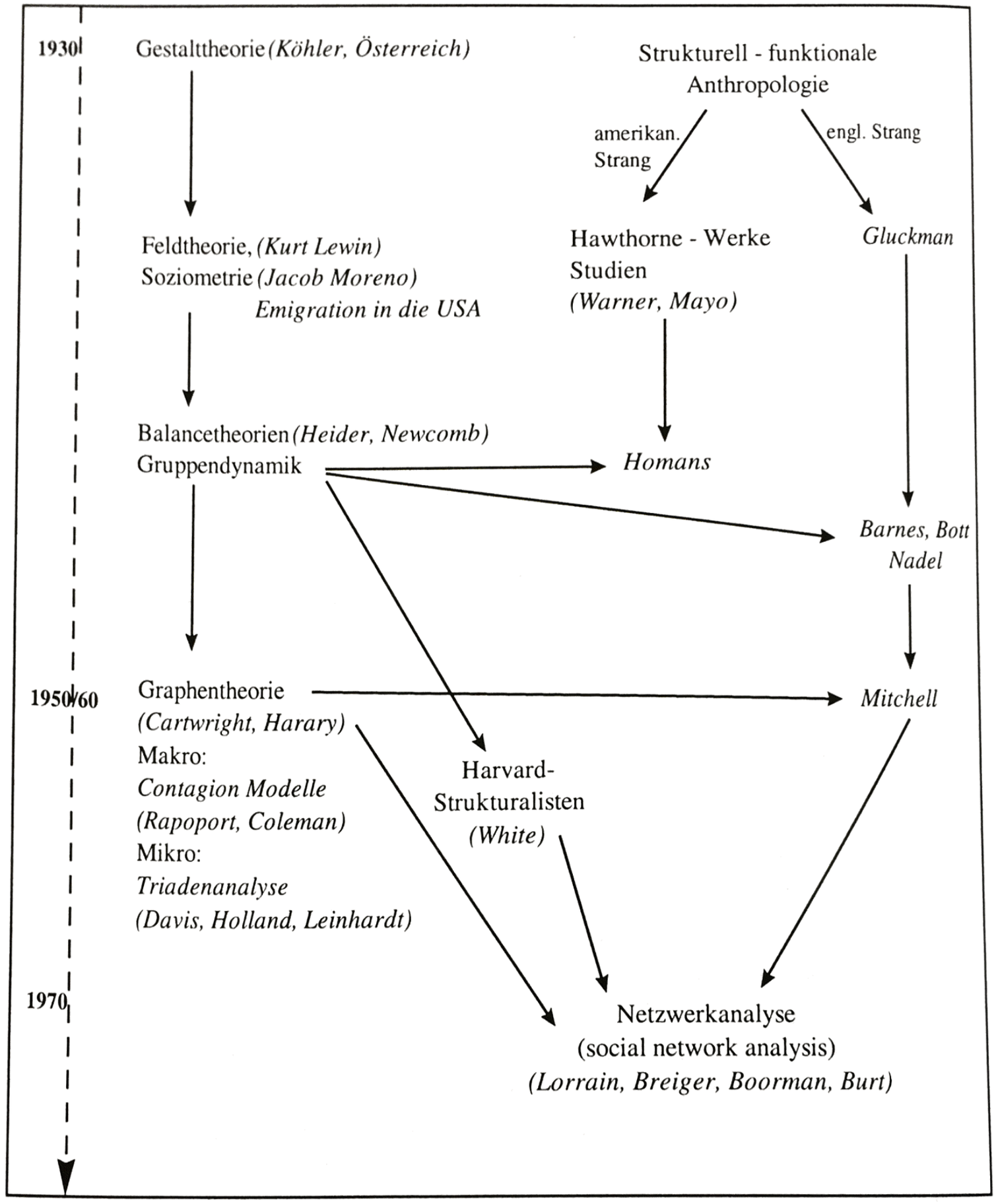
Recap

- We learnt about Network Science and Web Science and the differences in terms of the scope in each of the disciplines.

- We discussed many different examples of networks, mainly visualizations. Why visualizations - they are great for communicating insights.

- We understood the course requirements and know about the organization of the next three weeks.

Historical perspective on network science



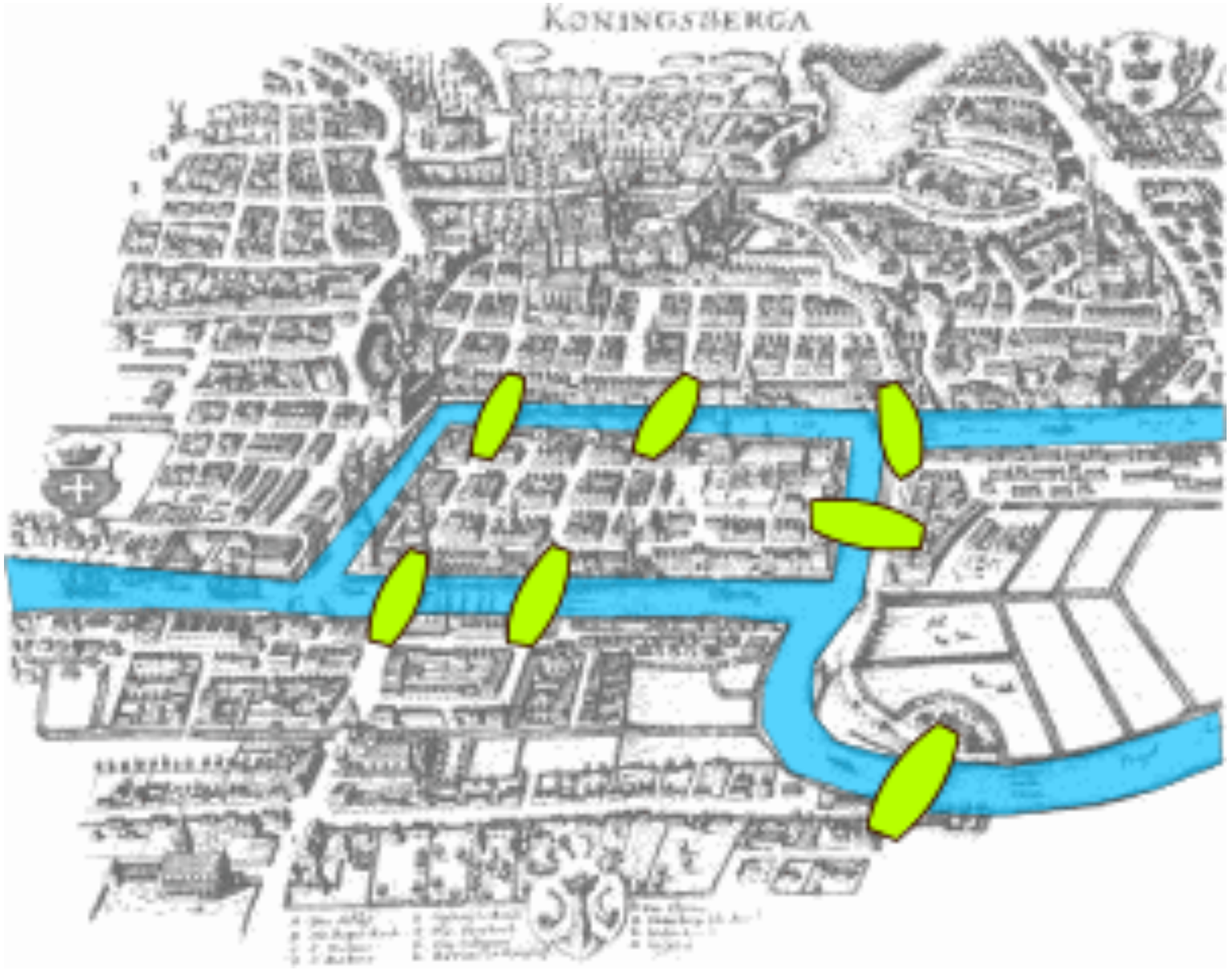
Jansen, Dorothea. Einführung in die Netzwerkanalyse: Grundlagen, Methoden, Forschungsbeispiele. Springer-Verlag, 2006.

Using graph theory for groups

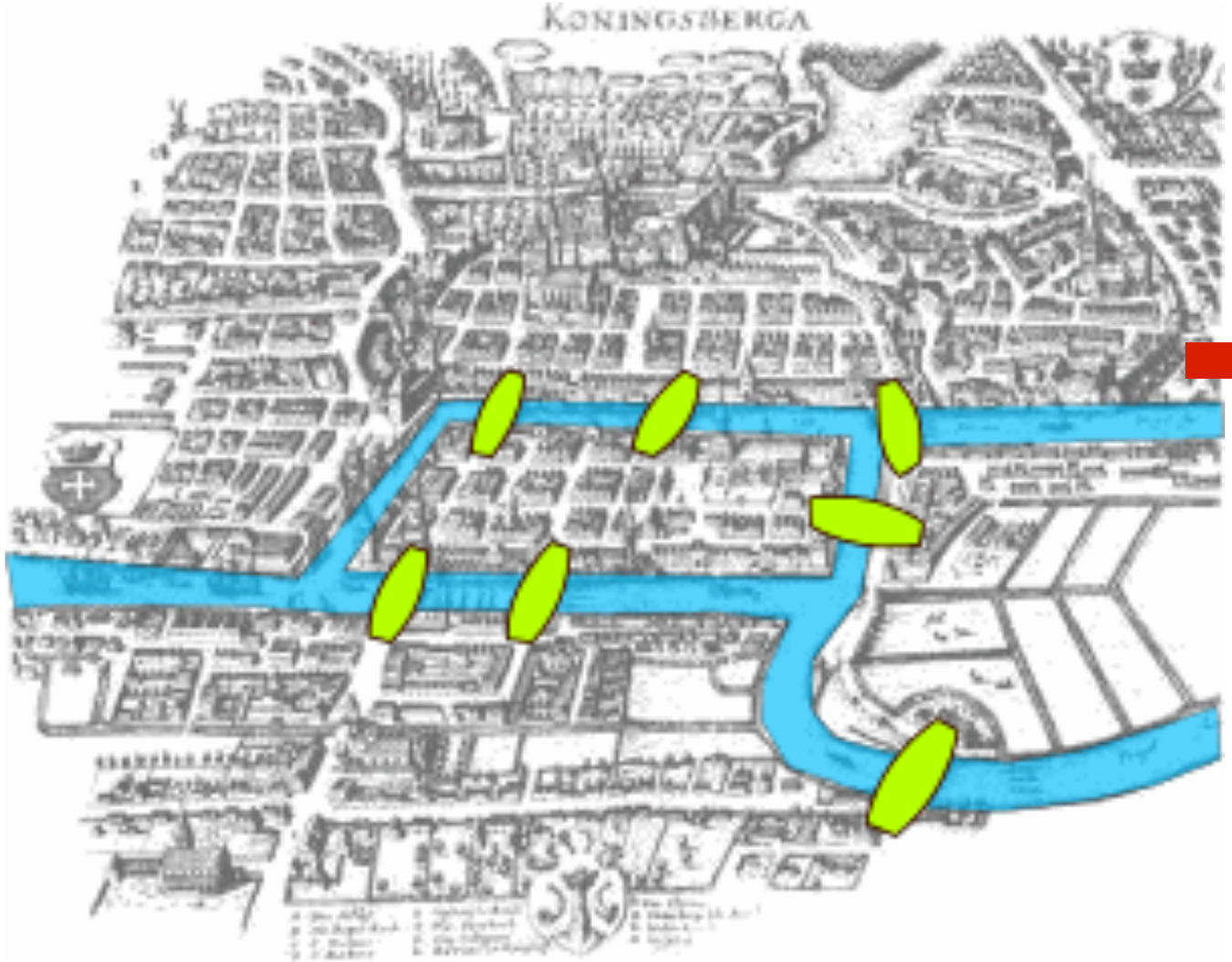
- In the 60er's, Cartwright, who, together with the mathematician Harary, pioneered the application of graph theory to group behavior
- Graph theory had first been formulated by König in 1936
- In 1950, König's book was republished in the United States and its ideas were developed in the work of Harary and Norman (1953)
- These mathematical ideas made possible a crucial breakthrough in the theory of group dynamics
- This breakthrough consisted of moving from the concept of cognitive balance in individual minds ("The enemy of my enemy is my friend") to that of interpersonal balance in groups

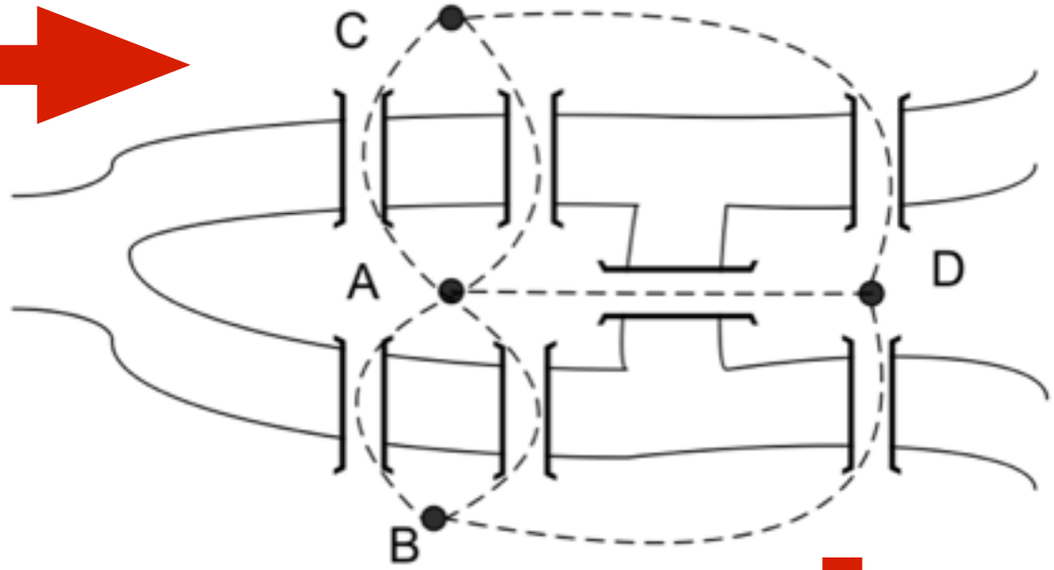
Jansen, Dorothea. Einführung in die Netzwerkanalyse: Grundlagen, Methoden, Forschungsbeispiele. Springer-Verlag, 2006.

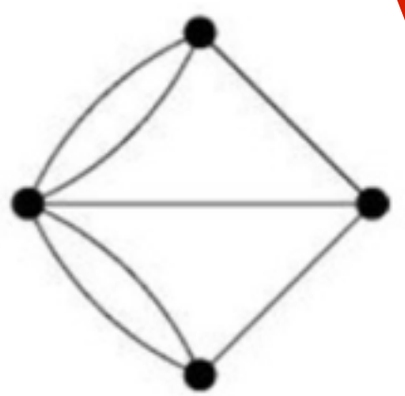
Seven Bridges of Königsberg (1736)



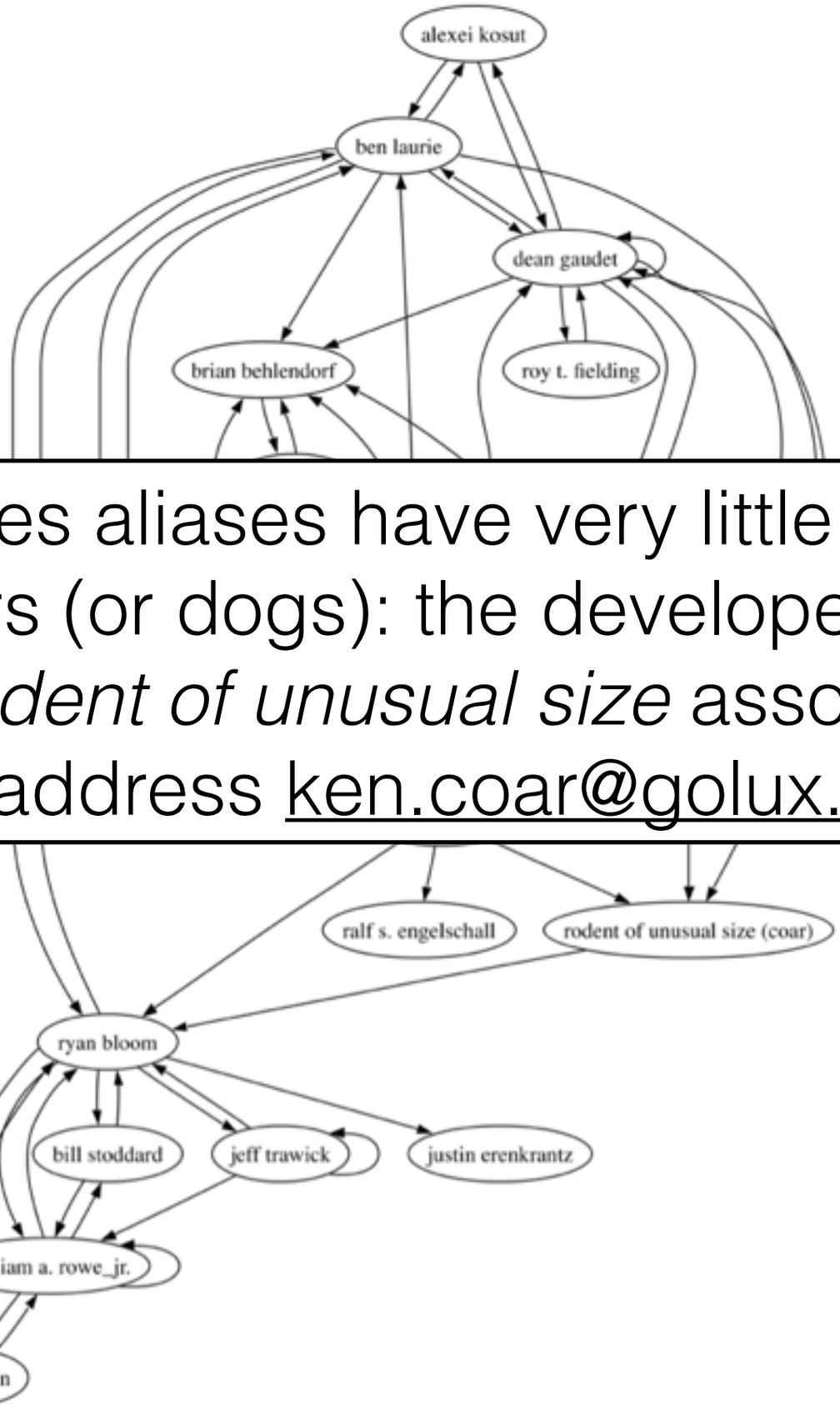
Seven Bridges of Königsberg (1736)







Who or what is “rodent of unusual size”?



Sometimes aliases have very little relationships to developers (or dogs): the developer Ken Coar uses the name *Rodent of unusual size* associated with email address ken.coar@golux.com

Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, and Anand Swaminathan. 2006. Mining email social networks. In Proceedings of the 2006 international workshop on Mining software repositories (MSR '06). ACM, New York, NY, USA, 137-143.

Today's outline

- Basic terminology
- Representing networks
- Visualizing networks
- Collecting networks
  - Social Production
  - Social Communication

Basic terminology

Network vs. graph

- A network refers to an informal concept describing an object composed of elements and interactions or connections between these elements

- A graph G = (V,E) is an abstract object formed by a set V of vertices (nodes) and a set E of edges (links) that join (connect) pairs of vertices

- The vertex set and edge set of a graph G are denoted by V(G) and E(G), respectively

- The cardinality of V is usually denoted by n, the cardinality of E by m

- If two vertices are joined by an edge, they are adjacent and we call them neighbors

Brandes, Ulrik, and Thomas Erlebach. Network analysis: methodological foundations. Vol. 3418. Springer Science & Business Media, 2005.

Undirected or directed graphs

- An undirected edge joining vertices u, v ∈ V is denoted by {u, v}
- In directed graphs, each directed edge (arc) has an origin (tail) and a destination (head)
- An edge with origin u ∈ V and destination v ∈ V is represented by an ordered pair (u, v). As a shorthand notation, an edge {u, v} or (u, v) can also be denoted by uv
- In a directed graph, uv is short for (u, v), while in an undirected graph, uv and vu are the same and both stand for {u, v}.
- Graphs that can have directed edges as well as undirected edges are called mixed graphs

Brandes, Ulrik, and Thomas Erlebach. Network analysis: methodological foundations. Vol. 3418. Springer Science & Business Media, 2005.

Multigraphs

- In both undirected and directed graphs, we may allow the edge set E to contain the same edge several times
- If an edge occurs several times in E, the copies of that edge are called parallel edges
- Graphs with parallel edges are also called **multigraphs**
- A graph is called **simple**, if each of its edges is contained in E only once, i.e., if the graph does not have parallel edges
- An edge joining a vertex to itself, i.e., an edge whose endvertices are identical, is called a **loop**
- A graph is called loop-free if it has **no loops**. We will assume all graphs to be loop-free unless specified otherwise

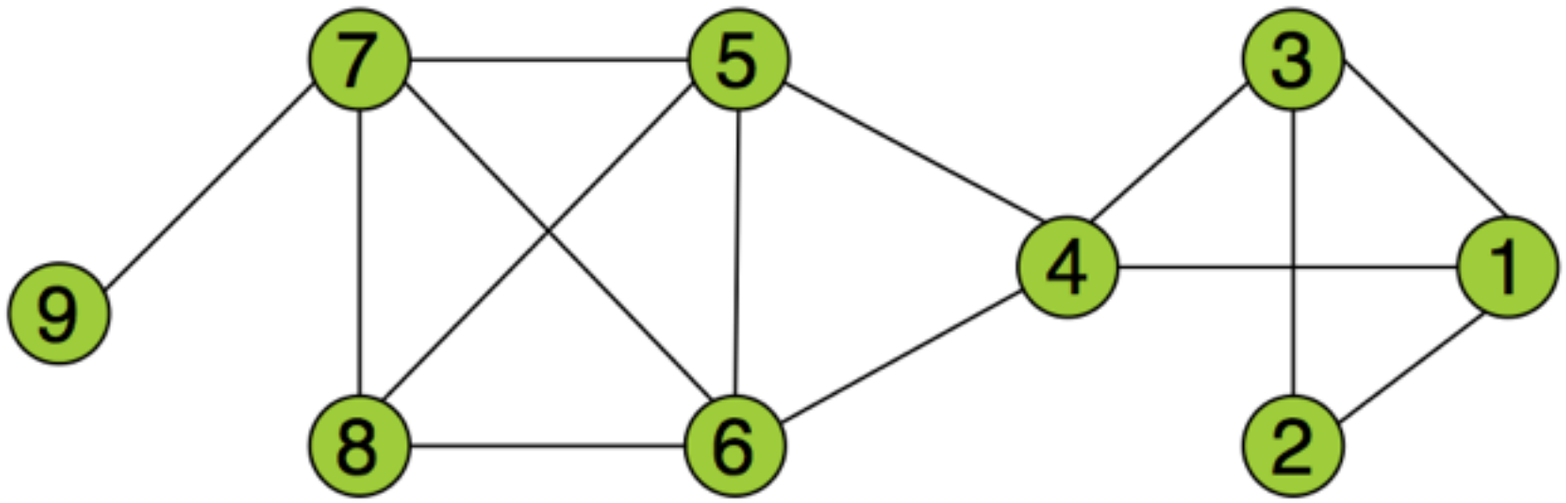
Brandes, Ulrik, and Thomas Erlebach. Network analysis: methodological foundations. Vol. 3418. Springer Science & Business Media, 2005.

Weighted graphs

- Often it is useful to associate numerical values (weights) with the edges
- Edge weights can be represented as a function $\omega : E \rightarrow R$ that assigns each edge $e \in E$ a weight $\omega(e)$
- Depending on the context, edge weights can describe various properties such as cost (e.g. travel time or distance), capacity, strength of interaction
- For most purposes, an unweighted graph $G = (V,E)$ is equivalent to a weighted graph with unit edge weights $\omega(e) = 1$ for all $e \in E$

Representing networks

A undirected simple network



V:	set of vertices in the network
E:	set of edges in the network
n:	the number of vertices (n=|V|)
m:	the number of edges (m=|E|)
u:	a vertex
(u,v):	an edge between vertices u and v

Vertex and edge list

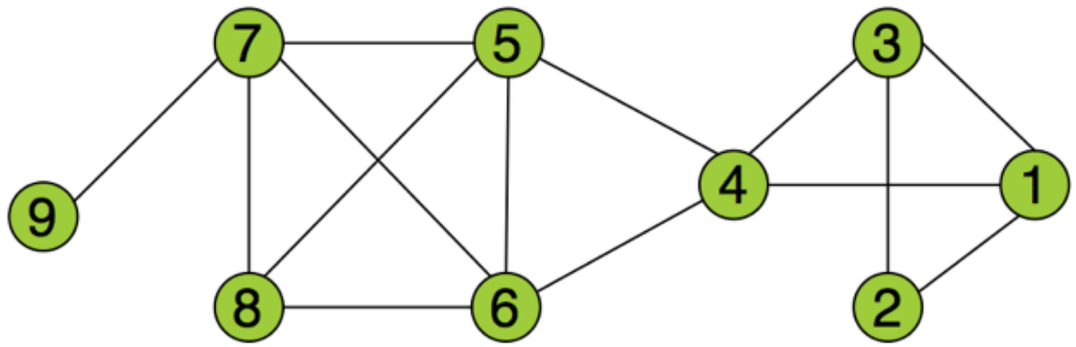
G = (V,E)

vertex list

1
2
3
4
5
6
7
8

edge list

1	2
1	3
1	4
2	3
3	4
4	5
4	6
5	6
5	7
5	8
6	7
6	8
7	8
7	9



How long does it take to find adjacent nodes?

How long does it take to check whether a given node is connected to another node or not?

Adjacency matrix

Adjacency matrix of a graph G=(V,E) consists of a |V|x|V| matrix A where;

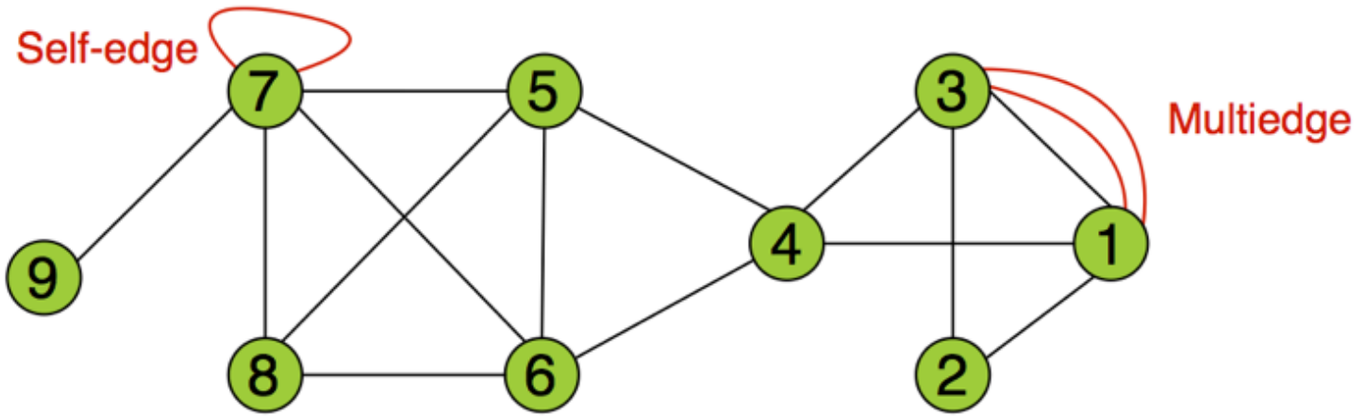
A_{uv} = 1 if G has an u-v edge.

A_{uv} = 0 otherwise.

	1	2	3	4	5	6	7	8	9
1	-	1	1	1	0	0	0	0	0
2	1	-	1	0	0	0	0	0	0
3	1	1	-	1	0	0	0	0	0
4	1	0	1	-	1	1	0	0	0
5	0	0	0	1	-	1	1	1	0
6	0	0	0	1	1	-	1	1	0
7	0	0	0	0	1	1	-	1	1
8	0	0	0	0	1	1	1	-	0
9	0	0	0	0	0	0	1	0	-

$$A = \begin{pmatrix} 011100000 \\ 101000000 \\ 110100000 \\ 101011000 \\ 000101110 \\ 000110110 \\ 000011011 \\ 000011100 \\ 000000100 \end{pmatrix}$$

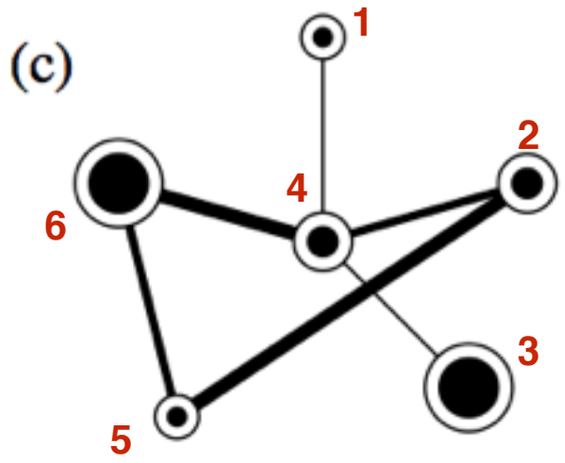
Adjacency matrix of a non-simple network



Node	1	2	3	4	5	6	7	8	9
1	-	1	3	1	0	0	0	0	0
2	1	-	1	0	0	0	0	0	0
3	3	1	-	1	0	0	0	0	0
4	1	0	1	-	1	1	0	0	0
5	0	0	0	1	-	1	1	1	0
6	0	0	0	1	1	-	1	1	0
7	0	0	0	0	1	1	2	1	1
8	0	0	0	0	1	1	1	-	0
9	0	0	0	0	0	0	1	0	-

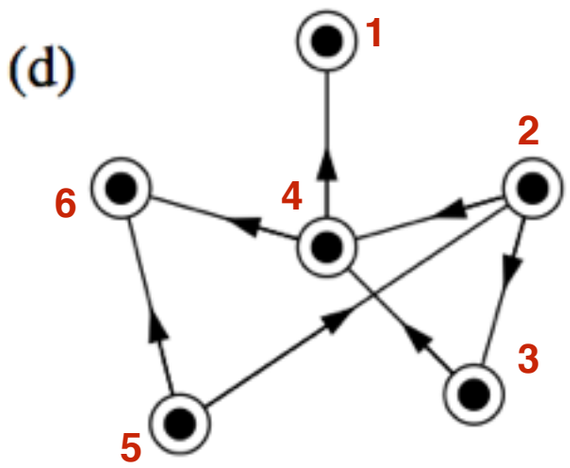
$$A = \begin{pmatrix} 013100000 \\ 101000000 \\ 310100000 \\ 101011000 \\ 000101110 \\ 000110110 \\ 000011211 \\ 000011100 \\ 000000100 \end{pmatrix}$$

Weighted and directed networks



$$A = \begin{pmatrix} 000100 \\ 000230 \\ 000100 \\ 121003 \\ 030002 \\ 000320 \end{pmatrix}$$

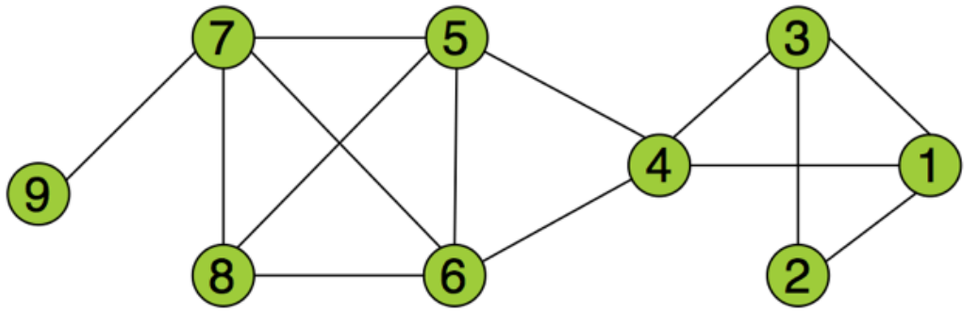
Please note:
I considered only the edge weights (thickest line = 3 and thinest line =1)

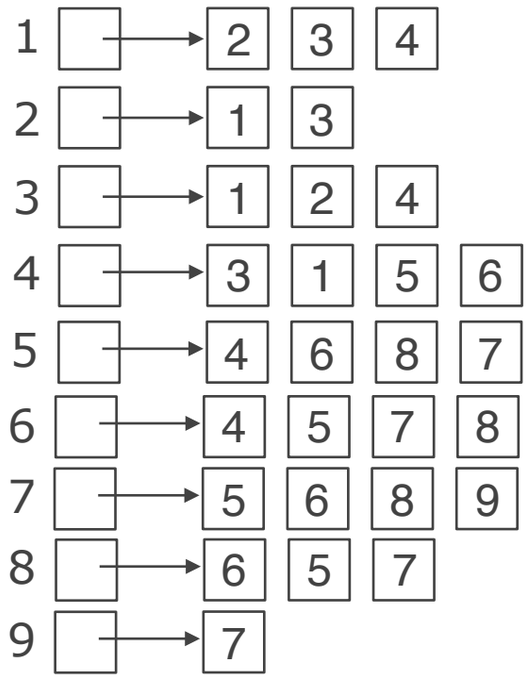


$$A = \begin{pmatrix} 000000 \\ 001100 \\ 000100 \\ 100001 \\ 010001 \\ 000000 \end{pmatrix}$$

Adjacency list

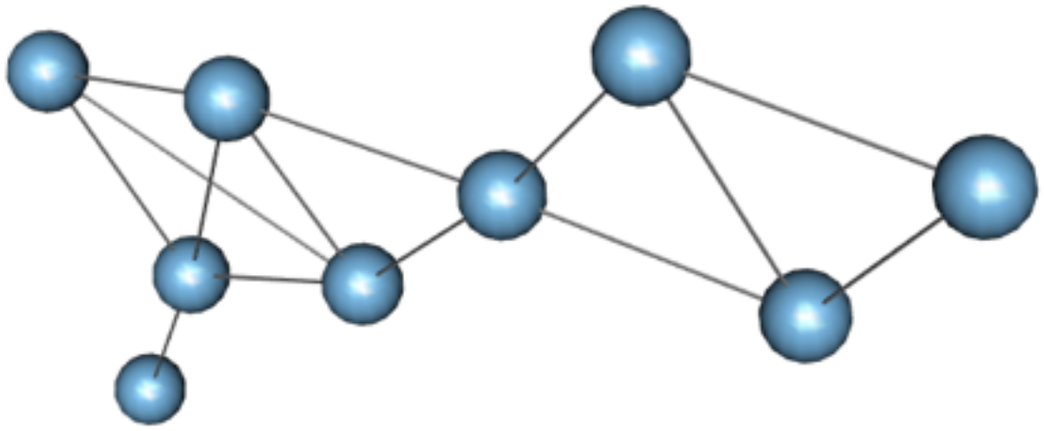
- Adjacency list representation of a graph G = (V , E) contains an array of vertices – lets call it *Adj*
- For each vertex u∈V, the adjacency list *Adj*[u] contains all adjacent vertices v such that there is an edge (u,v)∈E





Representations

```
> rglplot(g)
```



```
> get.adjacency(g)
9 x 9 sparse Matrix of class "dgCMatrix"

 [1,] . 1 1 1 . . . . .
 [2,] 1 . 1 . . . . . .
 [3,] 1 1 . 1 . . . . .
 [4,] 1 . 1 . 1 1 . . .
 [5,] . . . 1 . 1 1 1 .
 [6,] . . . 1 1 . 1 1 .
 [7,] . . . . 1 1 . 1 1
 [8,] . . . . 1 1 1 . .
 [9,] . . . . . . 1 . .
```

```
> get.edgelist(g)
      [,1] [,2]
 [1,]    1    2
 [2,]    1    3
 [3,]    1    4
 [4,]    2    3
 [5,]    3    4
 [6,]    4    5
 [7,]    4    6
 [8,]    5    6
 [9,]    5    7
[10,]    5    8
[11,]    6    7
[12,]    6    8
[13,]    7    8
[14,]    7    9
```

```
> get.adjlist(g)
[[1]]
[1] 2 3 4

[[2]]
[1] 1 3

[[3]]
[1] 1 2 4

[[4]]
[1] 1 3 5 6

[[5]]
[1] 4 6 7 8

[[6]]
[1] 4 5 7 8

[[7]]
[1] 5 6 8 9

[[8]]
[1] 5 6 7

[[9]]
[1] 7
```

Visualizing networks

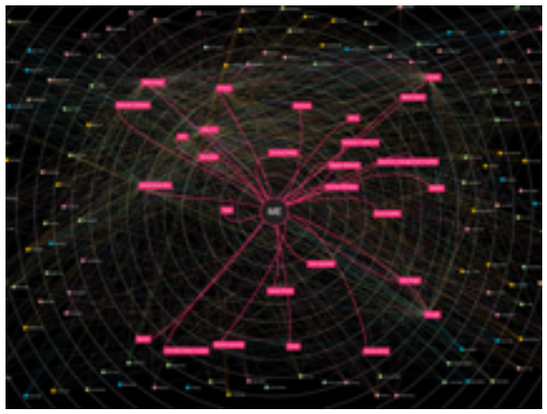
Network visualization

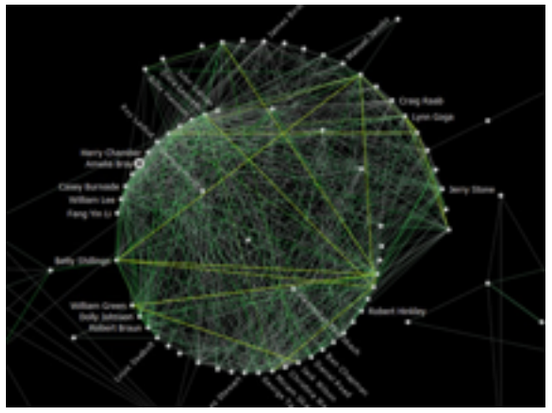
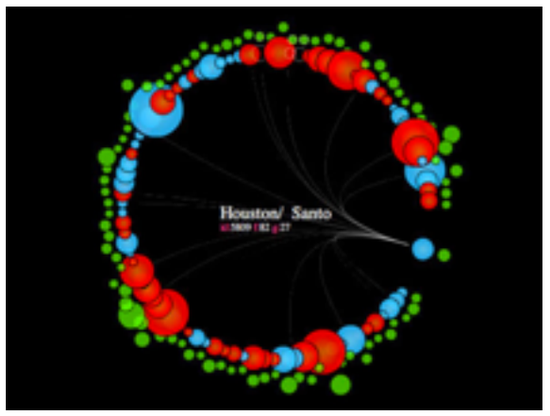
- Data visualization can be defined as any technique used to create images, diagrams, or animations in order to communicate a message.

- In general, data visualization is used as a way to aggregate large quantities of data and present them in a way that allows to
  - Quickly communicate rich messages (**communication**).
  - Discover new, previously unknown facts and relationships (**discovery**).
  - Get better insight into things we already know (**insight**).

- The visualization can be performed by using
  - network presentation that are made up of nodes and connection lines,
  - matrices where row and columns stand for actors or properties,
  - maps, and/or
  - a hybrid approach.

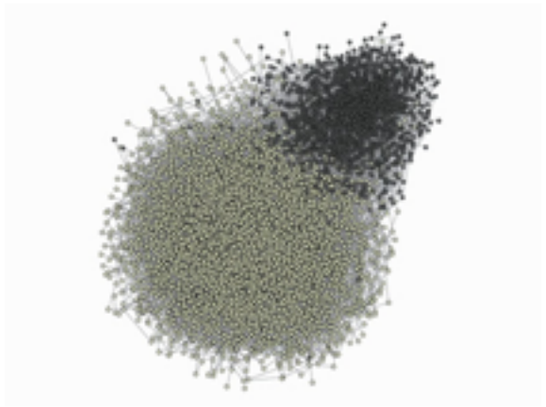
Abraham, Ajith, Aboul-Ella Hassanien, and Vaclav Snášel, eds. Computational social network analysis: Trends, tools and research advances. Springer Science & Business Media, 2009.

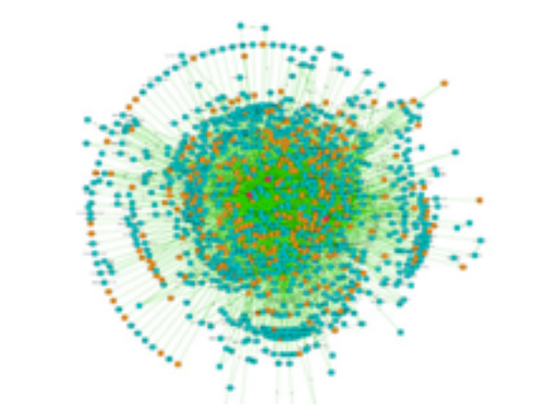
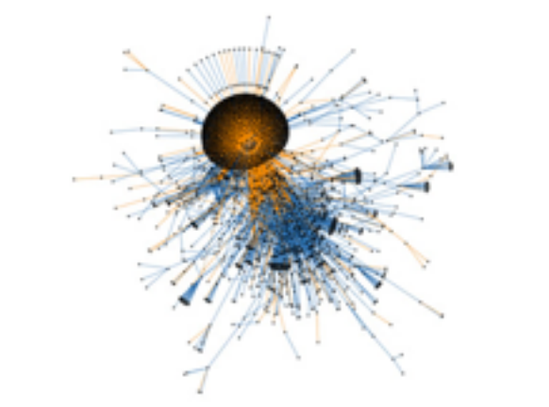
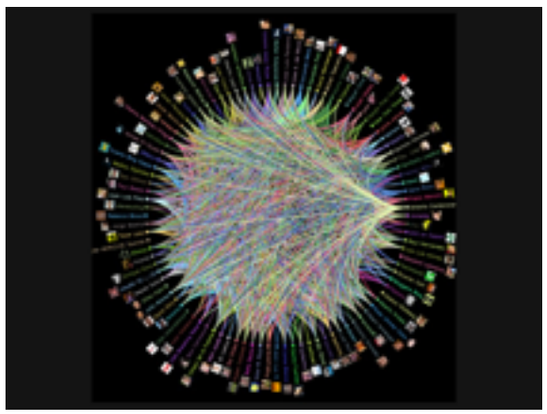
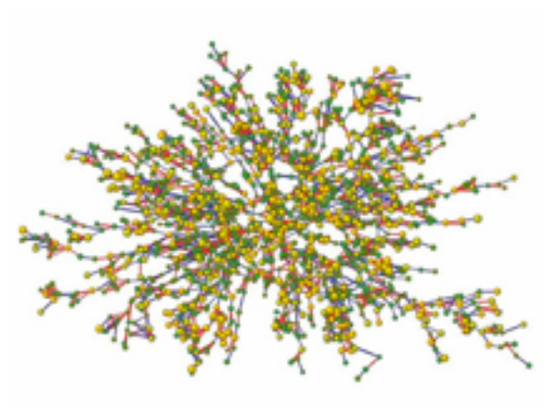
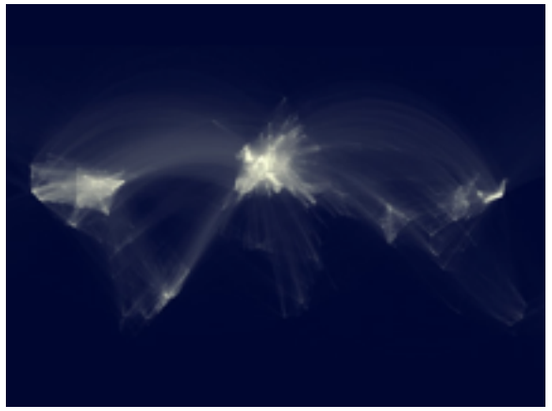
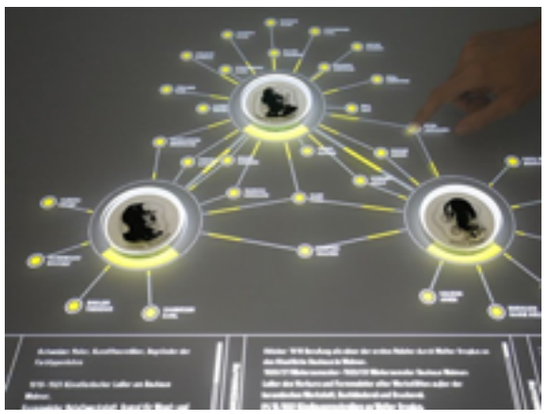


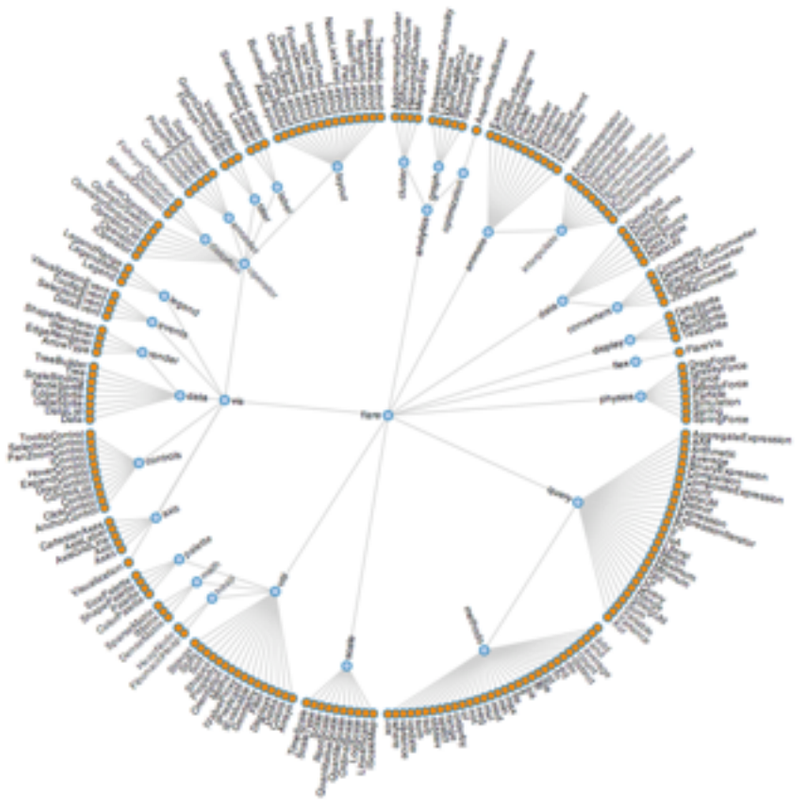


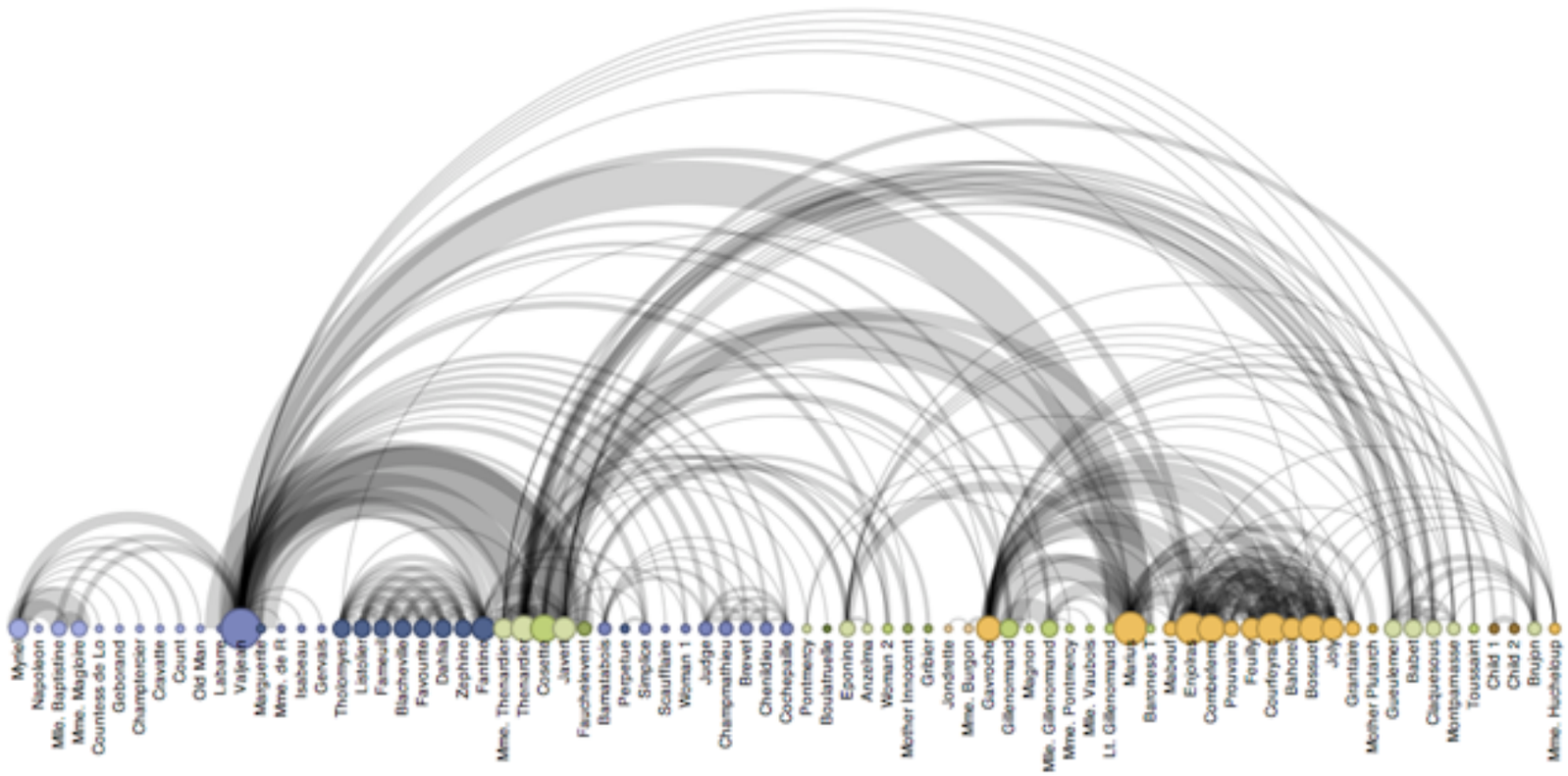


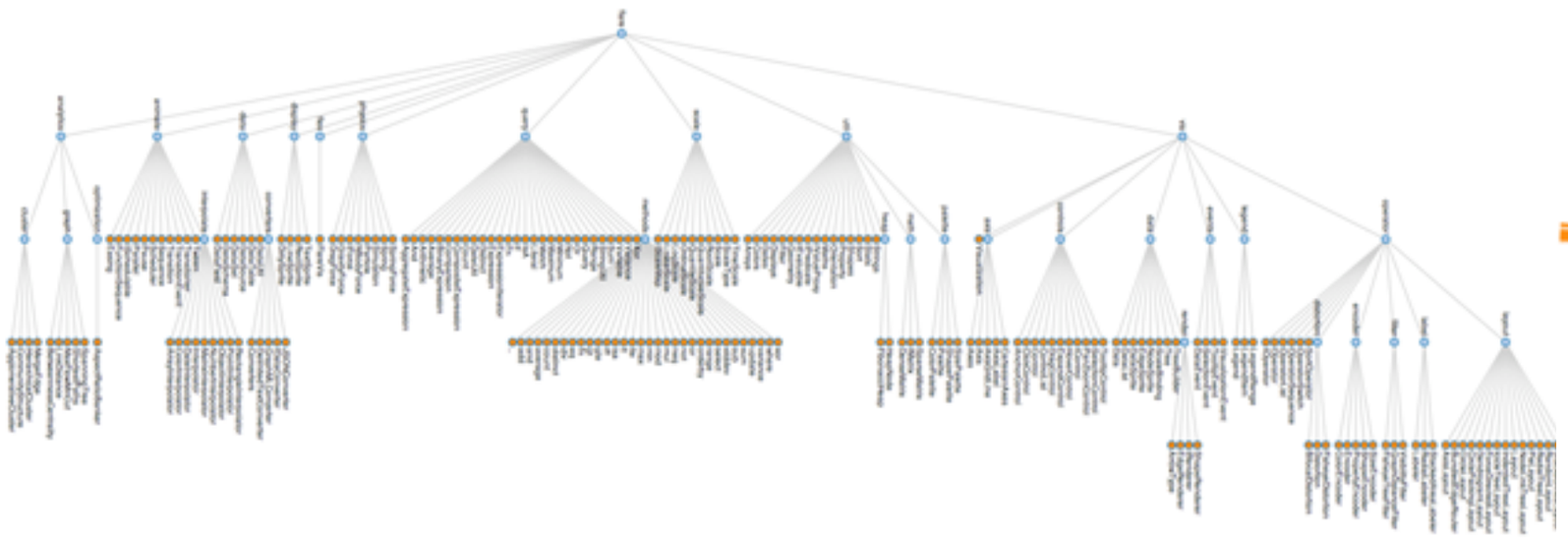


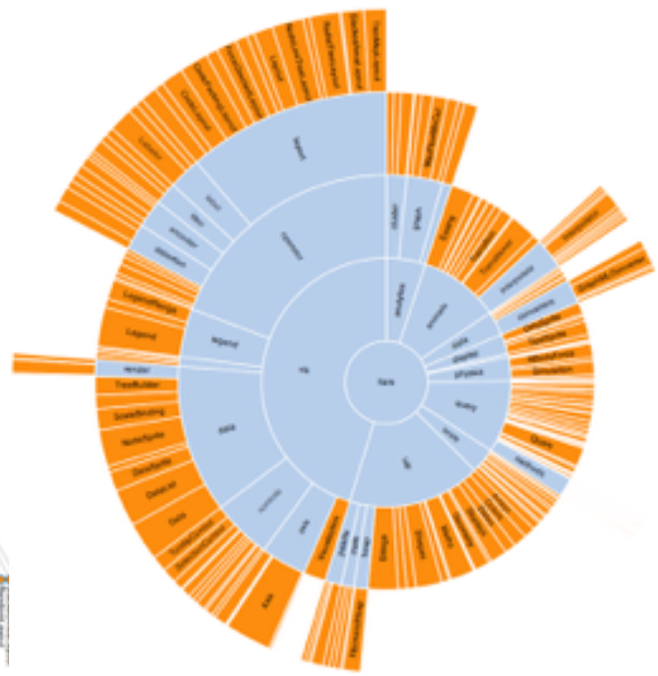







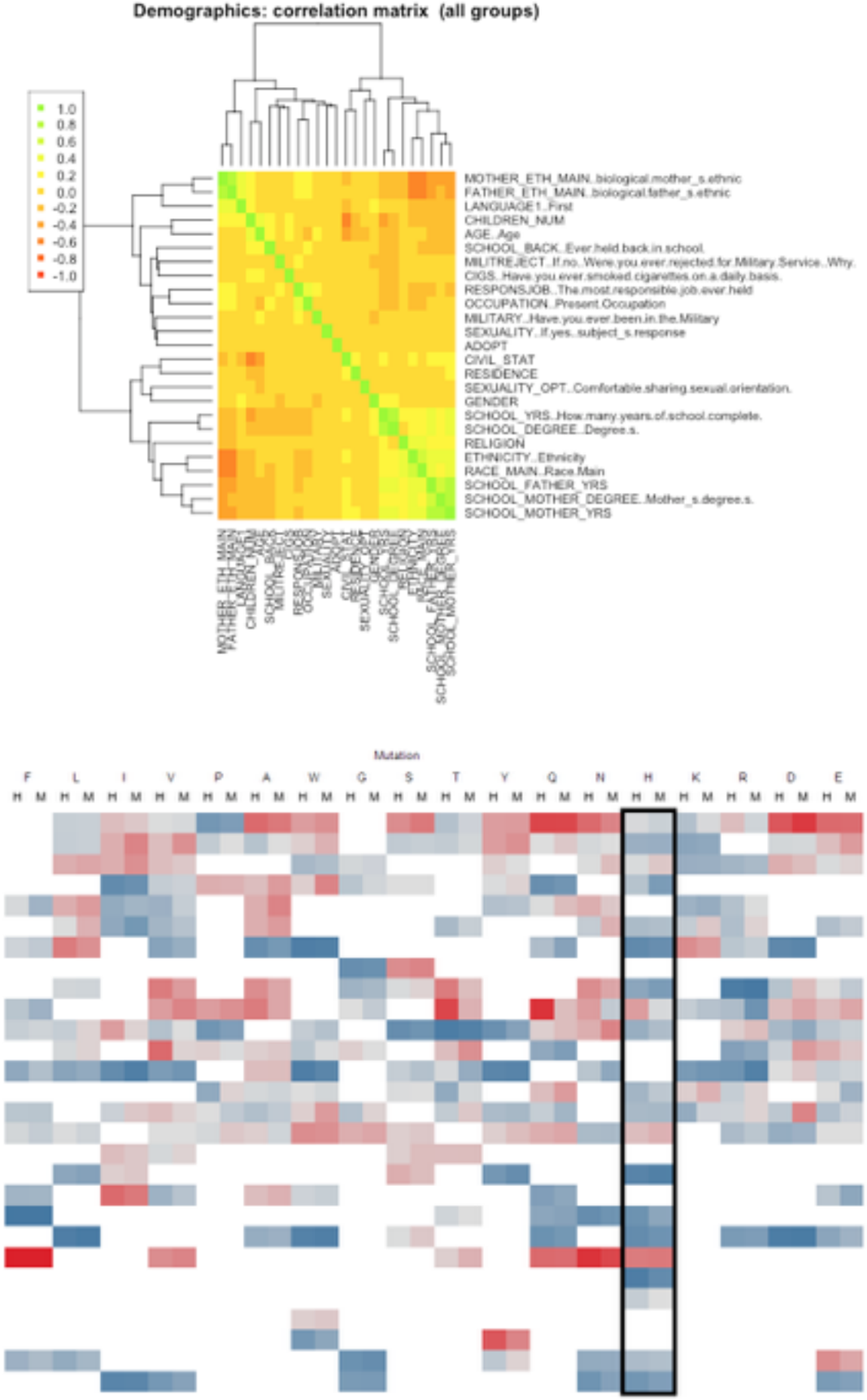
visualizations randomly chosen from visualcomplexity.org

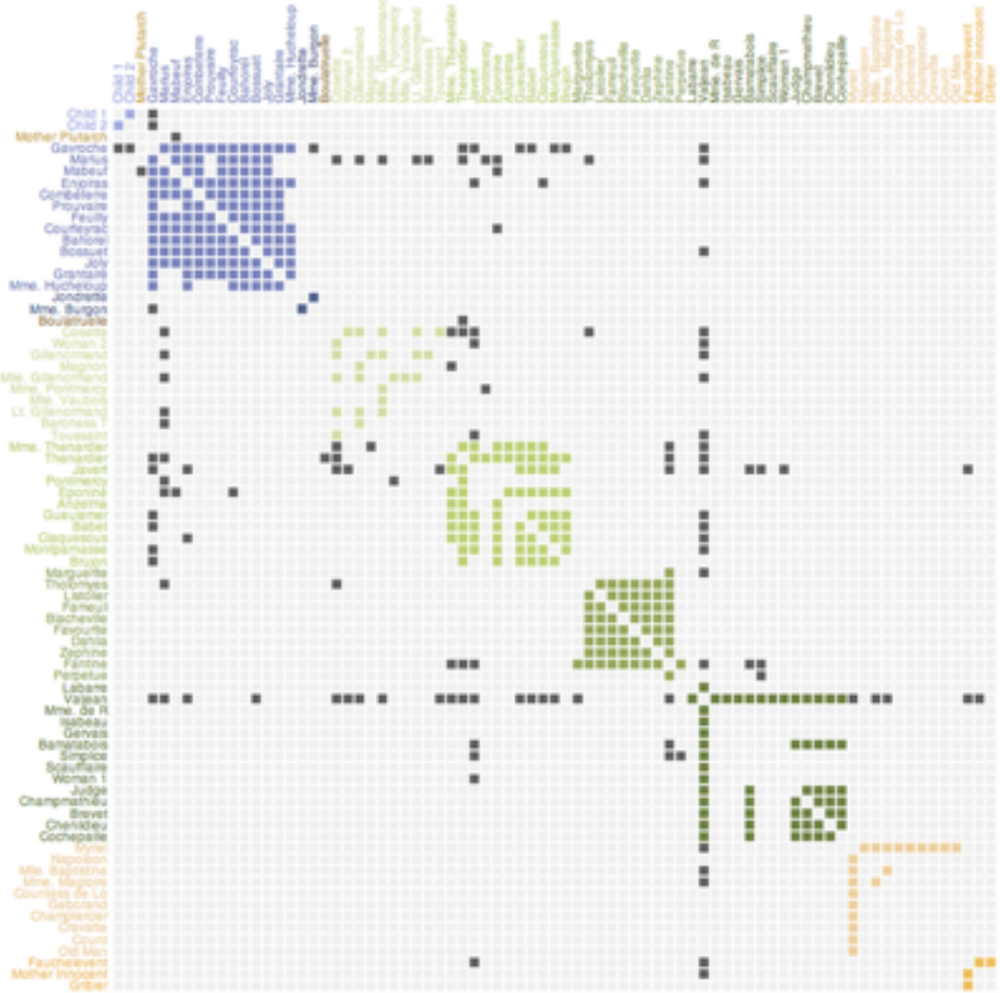



Myriel
Napoleon
Mlle. Baptistine
Mme. Magloire
Countess de Lo
Geborand
Champtercier
Cravatte
Count
Old Man
Labarre
Valjean
Marguerite
Mme. de R
Isabeau
Gervais
Tholomyes
Listolier
Fameuil
Blacheville
Favourite
Dahlia
Zephine
Fantine
Mme. Thenardier
Thenardier
Cosette
Javert
Fauchelevent
Bamatabois
Perpetue
Simplice
Scaufflaire
Woman 1
Judge
Champmathieu
Brevet
Chenildieu
Cochepaille
Pontmercy
Boulatruelle
Eponine
Anzelma
Woman 2
Mother Innocent
Gribier
Jondrette
Mme. Burgon
Gavroche
Gillenormand
Magnon
Mlle. Gillenormand
Mme. Pontmercy
Mlle. Vaubois
Lt. Gillenormand
Marius
Baroness T
Mabeuf
Enjolras
Combeferre
Prouvaire
Feuilly
Courfeyrac
Bahorel
Bossuet
Joly
Grantaire
Mother Plutarch
Gueulemer
Babet
Claquesous
Montparnasse
Toussaint
Child 1
Child 2
Brujon
Mme. Hucheloup

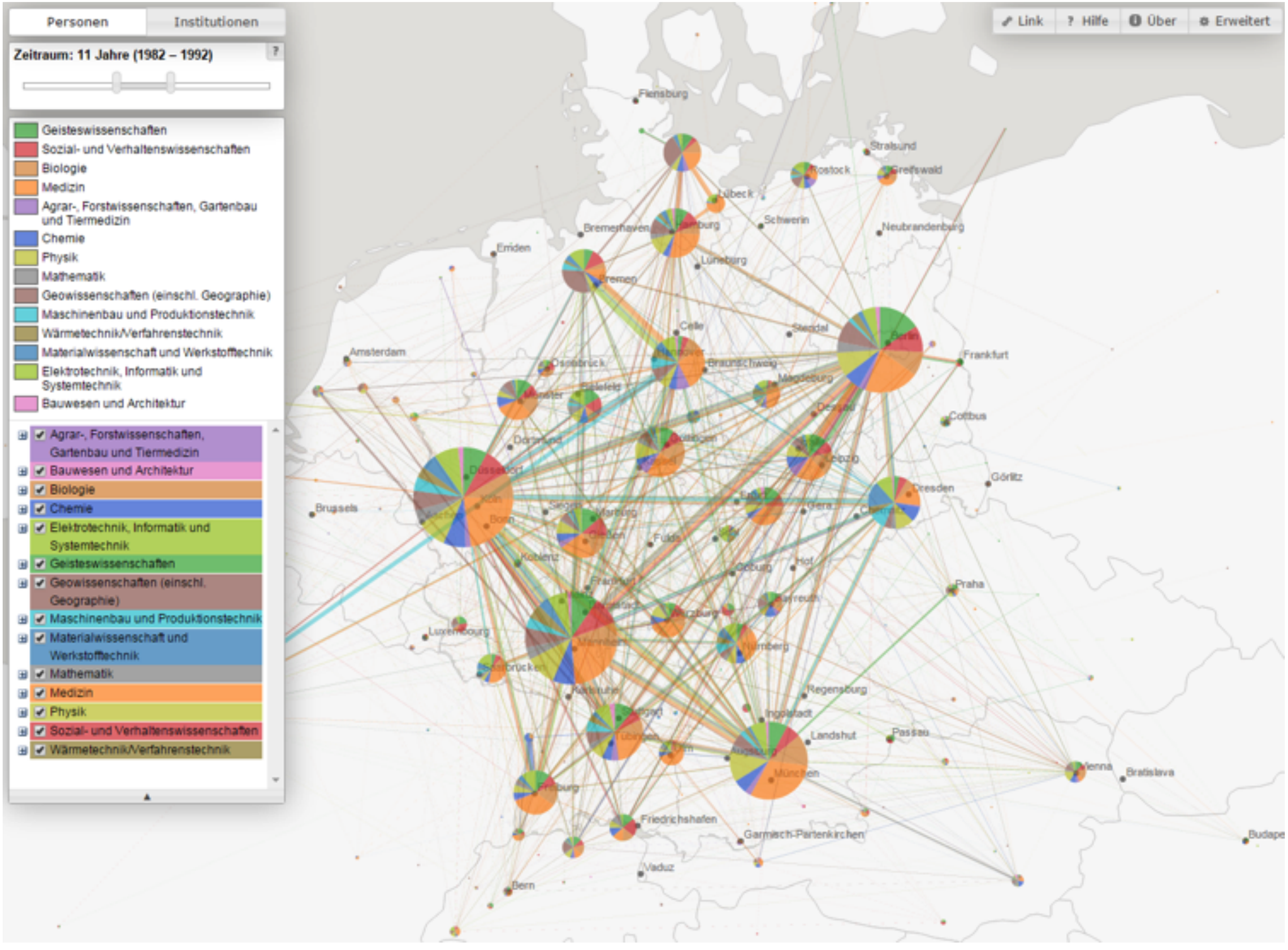





Demographics: correlation matrix (all groups)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
MOTHER_ETH_MAIN..biological.mother_s.ethnic
FATHER_ETH_MAIN..biological.father_s.ethnic
LANGUAGE1..First
CHILDREN_NUM
AGE..Age
SCHOOL_BACK..Ever.held.back.in.school.
MILITREJECT..If.no..Were.you.ever.rejected.for.Military.Service..Why.
CIGS..Have.you.ever.smoked.cigarettes.on.a.daily.basis.
RESPONSJOB..The.most.responsible.job.ever.held
OCCUPATION..Present.Occupation
MILITARY..Have.you.ever.been.in.the.Military
SEXUALITY..If.yes..subject_s.response
ADOPT
CIVIL_STAT
RESIDENCE
SEXUALITY_OPT..Comfortable.sharing.sexual.orientation.
GENDER
SCHOOL_YRS..How.many.years.of.school.complete.
SCHOOL_DEGREE..Degree.s.
RELIGION
ETHNICITY..Ethnicity
RACE_MAIN..Race.Main
SCHOOL_FATHER_YRS
SCHOOL_MOTHER_DEGREE..Mother_s.degree.s.
SCHOOL_MOTHER_YRS


Child 1
Child 2
Mother Plutarch
Gavroche
Marius
Mabeuf
Enjolras
Combeferre
Prouvaire
Feuilly
Courfeyrac
Bahorel
Bossuet
Joly
Grantaire
Mme. Hucheloup
Jondrette
Mme. Burgon
Boulatruelle
Unterlo
Women 2
Gillenormand
Magnon
Mlle. Gillenormand
Mme. Pontmercy
Mlle. Vaubois
Lt. Gillenormand
Baroness T
Tholomyes
Mme. Thenardier
Thenardier
Javert
Pontmercy
Eponina
Anzelma
Gueulemer
Babet
Claquesous
Montparnasse
Brujon
Marguerite
Tholomyes
Listolier
Fameuil
Blachevelle
Favourite
Dahlia
Zephine
Fantine
Perpetue
Labarre
Valjean
Mme. de R
Isabeau
Gervais
Barriabos
Simplice
Scaufflaire
Woman 1
Judge
Champmathieu
Brevet
Cherikdieu
Cochepaille
Myriel
Napoleon
Mlle. Baptistine
Mme. Magloire
Countess de Lo
Geborand
Champtercier
Cravatte
Count
Old Man
Fauchelevent
Mother Innocent
Gribier

Mutation
F L I V P A W G S T Y Q N H K R D E
H M
27 (G)
28 (F)
29 (T)
30 (F)
33 (S)
34 (S)
35 (Y)
36 (A)
56 (I)
57 (S)
58 (G)
59 (S)
61 (G)
62 (G)
63 (S)
64 (T)
105 (A)
106 (K)
108 (R)
109 (V)
110 (G)
111 (A)
112 (A)
113 (N)
114 (G)
116 (F)
117 (D)
118 (P)
IMGT position (with WT)
Colors
Max
1.00
Min


Personen
Institutionen
Zeitraum: 11 Jahre (1982 – 1992)
Geisteswissenschaften
Sozial- und Verhaltenswissenschaften
Biologie
Medizin
Agrar-, Forstwissenschaften, Gartenbau und Tiermedizin
Chemie
Physik
Mathematik
Geowissenschaften (einschl. Geographie)
Maschinenbau und Produktionstechnik
Wärmetechnik/Verfahrenstechnik
Materialwissenschaft und Werkstofftechnik
Elektrotechnik, Informatik und Systemtechnik
Bauwesen und Architektur
Link
Hilfe
Über
Erweitert
Flensburg
Stralsund
Greifswald
Rostock
Lübeck
Schwerin
Neubrandenburg
Bremerhaven
Emden
Lüneburg
Bremen
Celle
Stendal
Amsterdam
Osnabrück
Braunschweig
Magdeburg
Frankfurt
Cottbus
Dortmund
Leipzig
Görlitz
Dresden
Brussels
Köln
Bonn
Siegen
Marburg
Gießen
Fulda
Jena
Koblenz
Coburg
Hof
Praha
Bayreuth
Würzburg
Luxembourg
Nürnberg
Regensburg
Ingolstadt
Landshut
Passau
Wien
Bratislava
Budapest
Friedrichshafen
Garmisch-Partenkirchen
Vaduz
Bern

Collecting networks

Examples of Social/Pre-digital Media Systems

Size of Consumer Population	Size of Producer Population: Small	Medium	Large
Small	Instant messaging Personal messaging (e.g., within Facebook) Video chat Phone call Face-to-face office meeting	Committee report to a decision maker Online survey Social networking friend feed Twitter homepage showing tweets of people you follow	Professional services reports for decision makers Personalized suggestions based on recommender systems
Medium	"Social" or family blog Profile page on community site or social network Departmental email list Tweet sent to followers Wall post on Facebook	Group blog on niche topic Internet relay chat room Internal department wiki Facebook group Niche YouTube channels Local markets (e.g., Craigslist)	Professional report for specialty group NASA clickworkers[1] Idea-generation sites (e.g., IdeaConnection[2])
Large	Popular blog, podcast, or webcast Message to large forum or email list Popular Twitter user's tweet Popular YouTube video Company web site Novel	News rating site (e.g., Digg) Wikipedia page Television program Popular discussion forum Online user-generated databases (e.g., IMDB) or marketplace (e.g., Threadless)	Large online marketplaces (e.g., eBay) Wikipedia encyclopedia YouTube video sharing Flickr photo sharing Popular massively multiplayer game

[1] *www.scienceofcollaboratories.org/resources/collab.php?317*
[2] *www.ideaconnection.com/crowdsourcing/procter-gamble-00007.html*

Hansen, Derek, Ben Shneiderman, and Marc A. Smith. Analyzing social media networks with NodeXL: Insights from a connected world. Morgan Kaufmann, 2010.

Relational Metadata

- Technologies that "**capture**" communities' relational meta-data (Pingback and trackback in interblog networks, blogrolls, data provenance)

- Technologies to "**tag**" communities' relational metadata (from Dublin Core taxonomies to folksonomies ('wisdom of crowds') like
  - Tagging photos (Flickr)
  - Tagging images (ESP)
  - Tagging blogs (Technorati)
  - Tagging news stories (digg)
  - Social bookmarking (del.icio.us)
  - Social citations (CiteULike.org)
  - Social libraries (discogs.com, LibraryThing.com)
  - Social shopping (SwagRoll, Kaboodle, thethingsiwant.com)
  - Social networks (FOAF, XFN, MySpace, Facebook)

- Technologies to "**manifest**" communities' relational metadata (Tagclouds, Recommender systems, Rating/Reputation systems)

Digital Harvesting

Text Mining

Web Crawling

API

Web of Science Citation

Adapted from Noshir Contractor: Web Science meets Network Science

Network data collection

- The objective of the data collection is to provide a data set that could help analyze the effects digital social networks have in the different aspects of social activities such as
  - (i) their generation and
  - (ii) their spatial distribution

- The generation of social activities considers both the members' (actors') propensity and opportunity to engage in social activities and their interactions with other members in the network

- The spatial distribution describes the activity spaces where members move around when they interact.

Abraham, Ajith, Aboul-Ella Hassanien, and Vaclav Snášel, eds. Computational social network analysis: Trends, tools and research advances. Springer Science & Business Media, 2009.

Methods for data collection

- *The Socio-centric method*
  - maps the relations among actors ‘that are regarded for analytical purposes as bounded social collectives’
  - most appropriate for tight-bound networks

- *The Ego-centric method*
  - maps the relations of a key individual
  - this is most appropriate for loose-bound networks
  - method focuses on the individual, rather than on the network as a whole
  - allows selecting focal actors (egos), and identifies the nodes to which they are connected to

Computational social network analysis: Trends, tools and research advances. Springer Science & Business Media, 2009.
Abraham, Ajith, Aboul-Ella Hassanien, and Vaclav Snášel, eds.

Example on email based communication networks

Social Production: Wikipedia

article | discussion | view source | history

Berlin

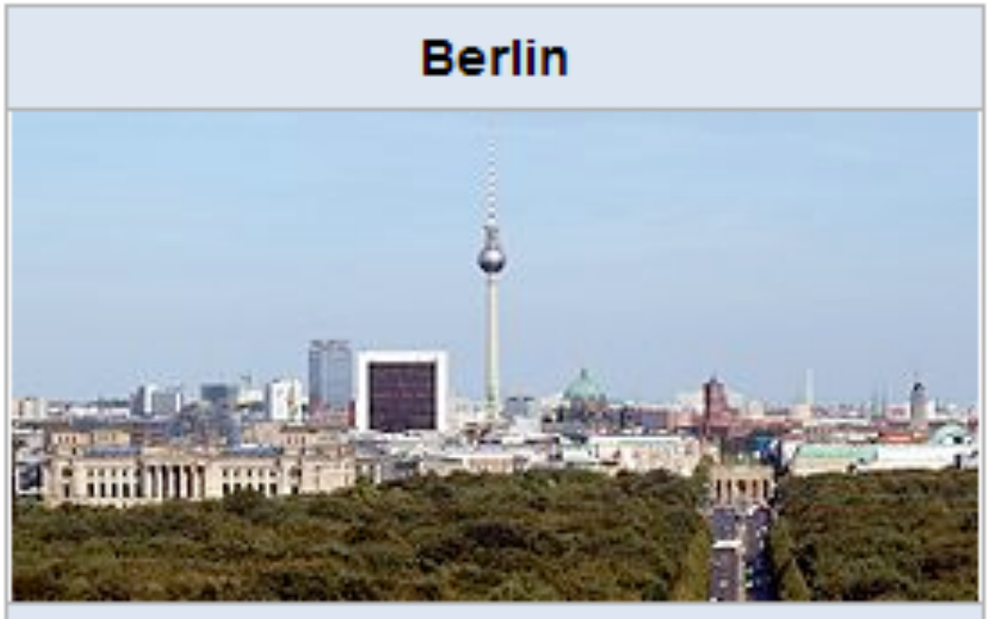
From Wikipedia, the free encyclopedia

Coordinates: 52.5°N 13.4°E

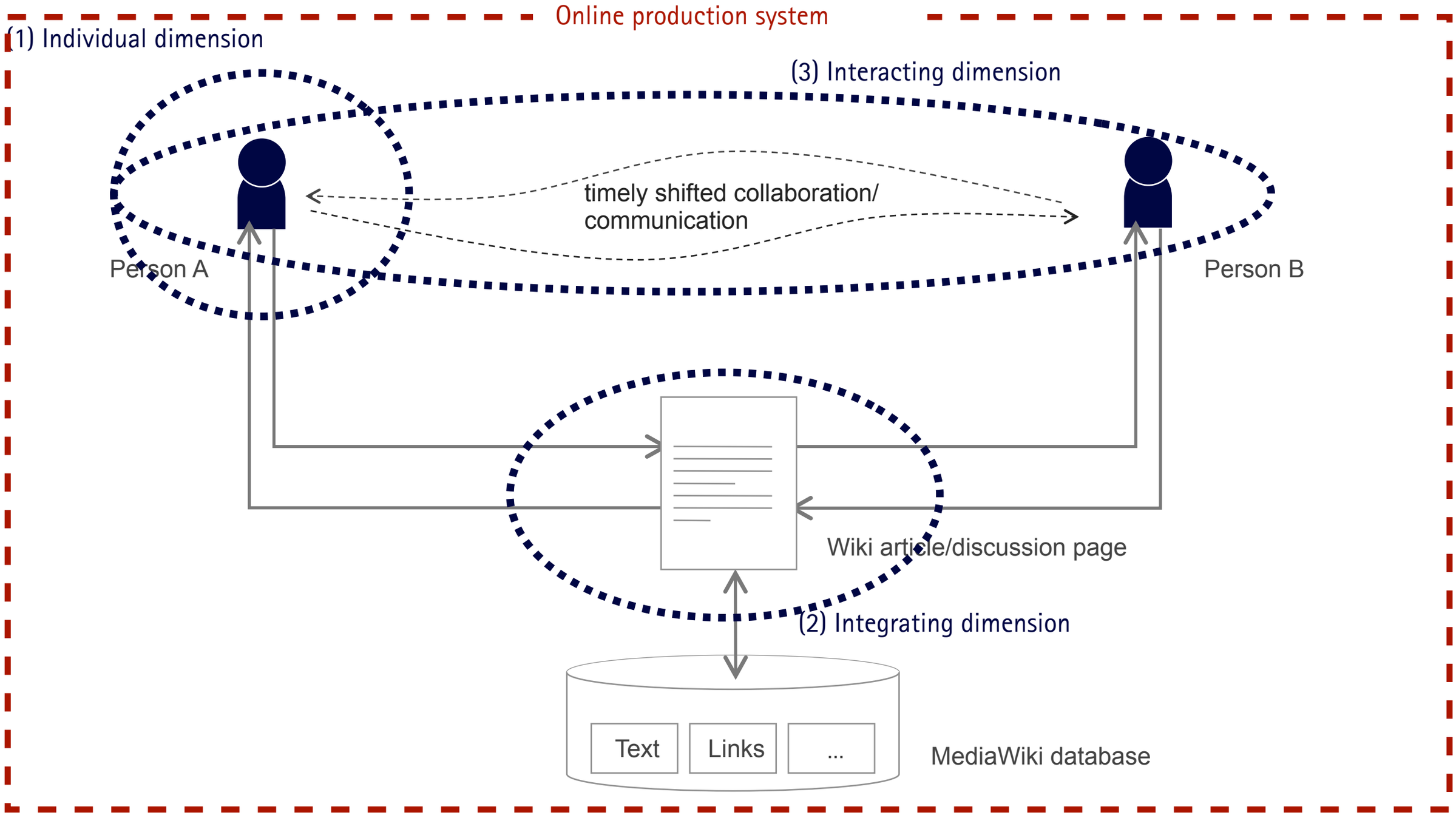
This article is about the capital of Germany. For other uses, see Berlin (disambiguation).

Berlin

(German pronunciation (help·info)) is the capital city and one of sixteen states of Germany. With a population of 3.4 million within its city limits, Berlin is Germany's largest city. It is the second most populous city and the eighth most populous urban area in the European Union.[2] Located in northeastern Germany, it is the center of the Berlin-Brandenburg metropolitan area, comprising 5 million people from over 190 nations.[3]



Wikis as social information space



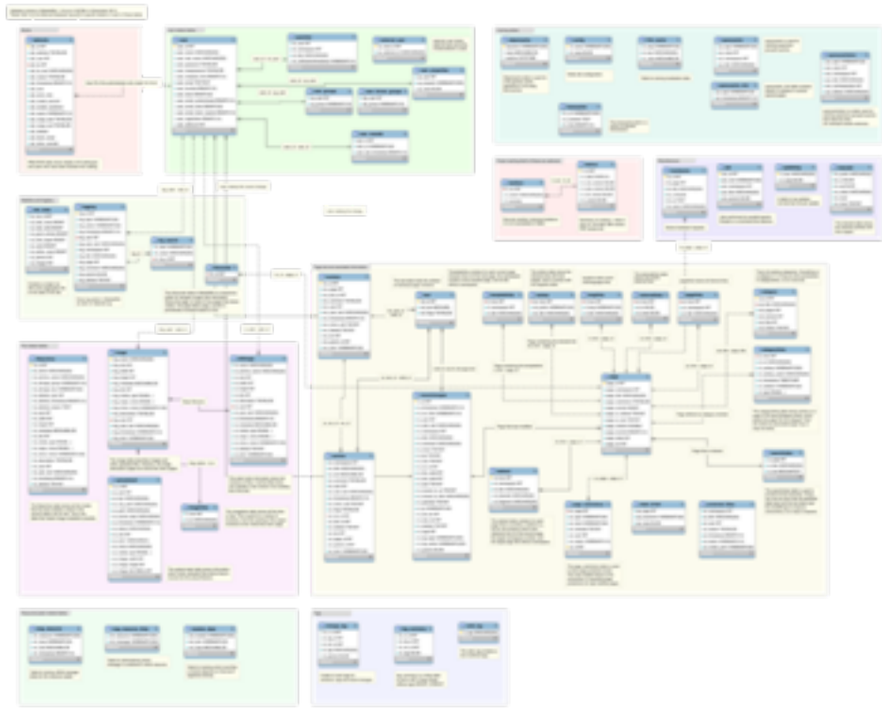
Possibilities of collecting data

API

- api.php is a web-service provided by MediaWiki since September 2006
- It has a multitude of different request-parameters as well as three different output-formats (HTML, XML and JSON)
- It can be accessed on most Wikipedias as well as many other MediaWiki-instances worldwide
- In case of the English Wikipedia, the base-url for the api.php is http://en.wikipedia.org/w/api.php.

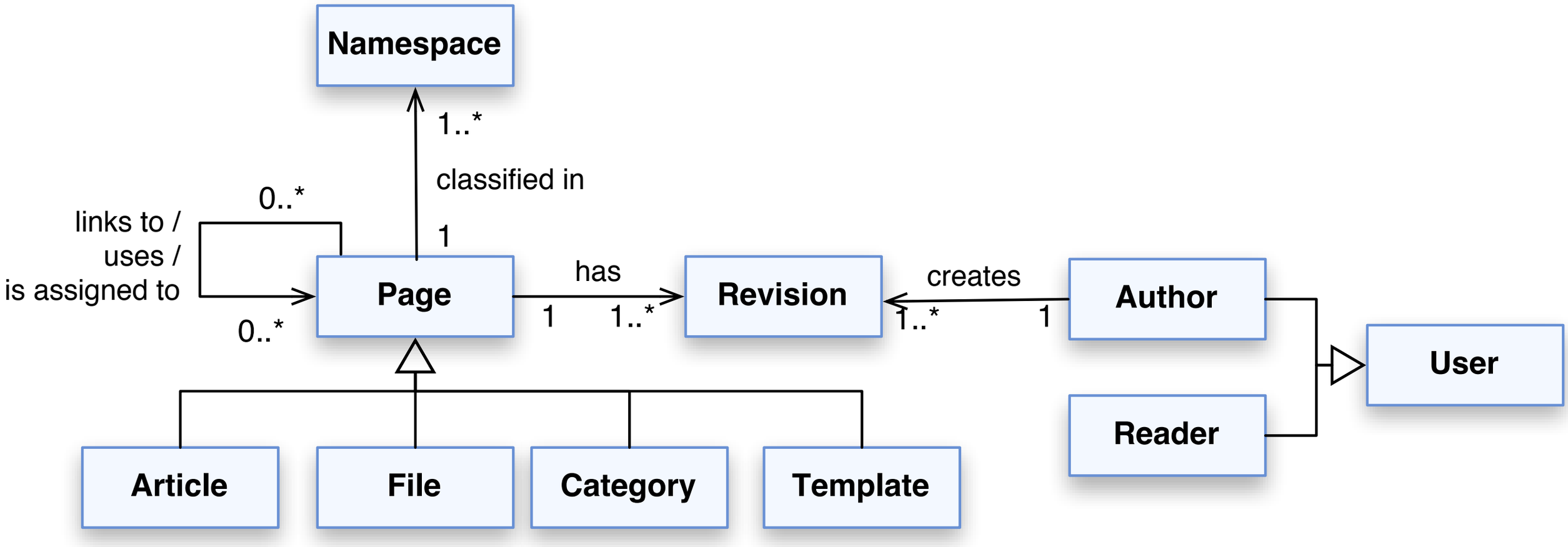
Database Dump

- The Wikimedia Foundation offers complete dumps of most Wikipedia tables
- These dumps can be found here: http://dumps.wikimedia.org/



http://www.mediawiki.org/wiki/Manual:Database_layout

Basic concepts in MediaWiki



Exemplary networks in Wikipedia

- *Collaboration/co-authorship/discussion/coordination network*

  $G_{coll} = (V_{authors}, E_{revision}, W_{\#jointedits})$,

  where $V_{authors}$ are authors, $E_{revision}$ are mutually edited articles and $W_{\#jointedits}$ are the number of co-located edits

  Depending of the used name space, the meaning of the network changes (ns:1 article, ns:2 discussion)

- *Wiki-link/article network*

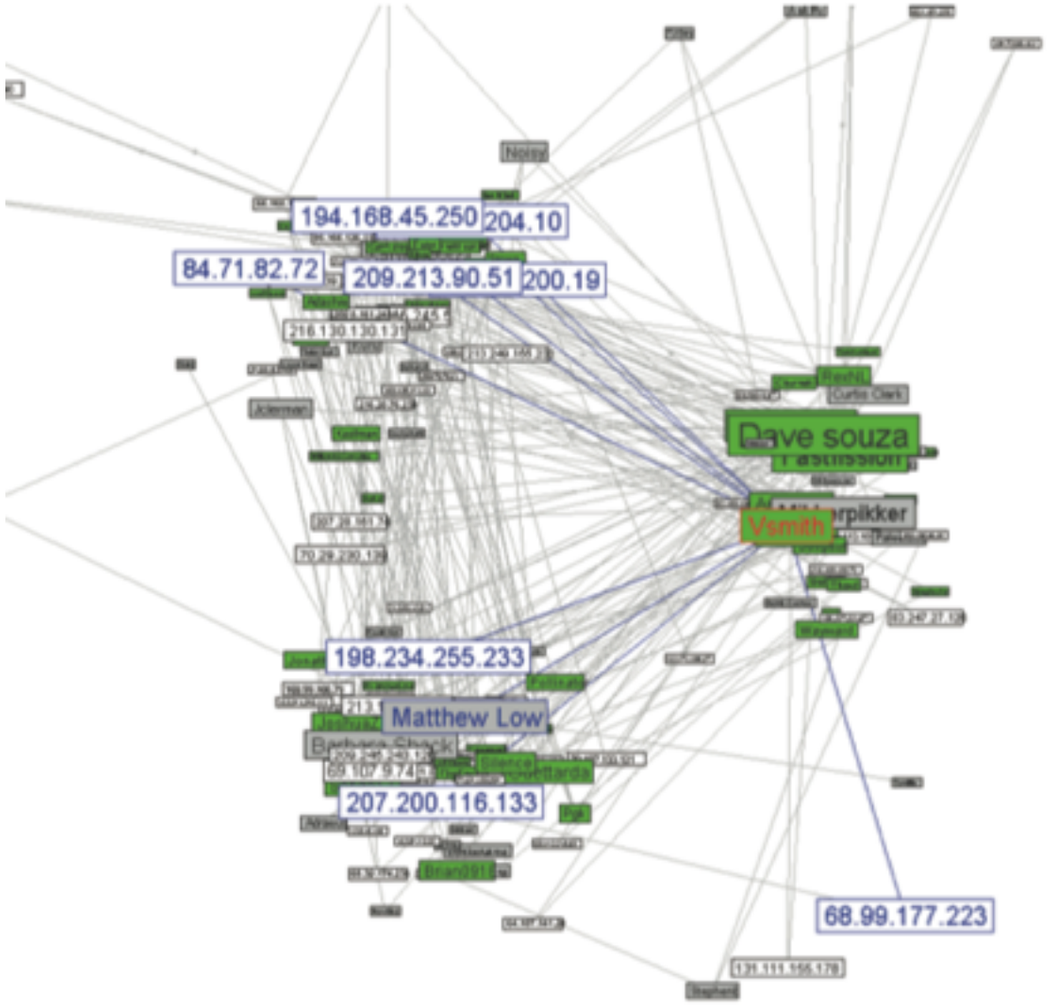
  $G_{link} = (V_{article}, E_{links})$

  where $V_{article}$ are the article in ns:0 and E_{links} are wiki-links between these pages

Revert Network

Suh, B., Chi, E. H., Pendleton, B. A., & Kittur, A. 2007, 'Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations', Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on, 163-170.

- A user conflict model based on users' editing histories, specifically revisions that void previous edits, known as "reverts"
- Model extracts reverts from Wikipedia editing history and composes a node-link diagram where a user is denoted as a node and a revert relationship as a link
- Based on this model, a tool called *Revert Graph* is developed that visualizes the revert relationships between opinion groups
- Revealed patterns such as: (a) formation of opinion groups; (b) patterns of mediation; (c) fighting of vandalism; (d) identification of major controversial users and topics



Color-coded based on users' registration status: An administrator is drawn as a *green* node, a normal registered user as a *grey* node, and an unregistered anonymous user as a *white* node

Social Communication: E-Mail

Email data capture

- There are a two strategies to capture email data:'server-side' and 'client-side' strategies

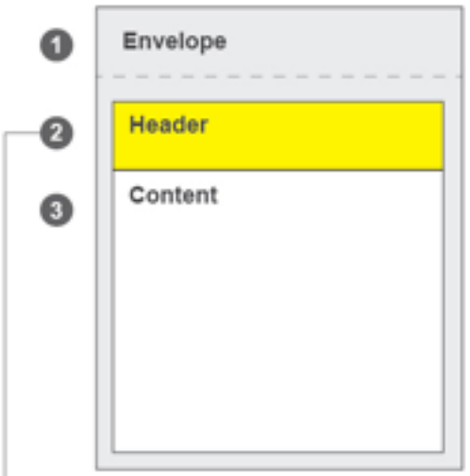
- Server-side:
  - If one captures the entire email spool for a university domain (such as @fu-berlin.de), one is assuming that this is the primary email for these individuals
  - Example: (Kossinets and Watts, 2006)

- Client-side:
  - Involves the use either of email monitoring software or parsing scripts
  - Is well suited to personal network analysis as one can capture the network on the client's computer and compare it to similarly captured networks
  - Is less than ideal for whole network analysis as one only has the mail that is seen by a particular address

Hogan, B. (2008) Analyzing Social Networks via the Internet. In: N.Fielding, R.Lee and G.Blank (eds) The Handbook of Online Research Methods. Thousand Oaks, CA: Sage, pp. 141-160.

E-Mail metadata

Email **Structure**



Email **Header**

Received: from relay.hackingteam.com (192.168.100.52) by EXCHANGE.hackingteam.local (192.168.100.51) with Microsoft SMTP Server id 14.3.123.3; Thu, 2 Apr 2015 21:52:40 +0200
Received: from mail.hackingteam.it (unknown [192.168.100.50]) by relay.hackingteam.com (Postfix) with ESMTP id A510360062; Thu, 2 Apr 2015 20:30:15 +0100 (BST)
Received: by mail.hackingteam.it (Postfix) id 789AAB6600B; Thu, 2 Apr 2015 21:52:40 +0200 (CEST)
Delivered-To: amministrazione@hackingteam.it
Received: from manta.hackingteam.com (manta.hackingteam.com [192.168.100.25]) by mail.hackingteam.it (Postfix) with ESMTP id 6EB0A2BC0DA for <amministrazione@hackingteam.it>; Thu, 2 Apr 2015 21:52:40 +0200 (CEST)
X-ASG-Debug-ID: 1428004358-056a757fe59b680001-hkhb8U
Received: from mail-wi0-f178.google.com (mail-wi0-f178.google.com [209.85.212.178]) by manta.hackingteam.com with ESMTP id FPwDcTRJ3dTkL9UM for <amministrazione@hackingteam.it>; Thu, 02 Apr 2015 21:52:38 +0200 (CEST)
X-Barracuda-Envelope-From: giovanni.cino@gmail.com
X-Barracuda-IPDD: Level1 [gmail.com/209.85.212.178]
X-Barracuda-Apparent-Source-IP: 209.85.212.178
Received: by wizk4 with SMTP id k4so26212015wiz.1 for <amministrazione@hackingteam.it>; Thu, 02 Apr 2015 12:52:38 -0700 (PDT)
DKIM-Signature: v=1; a=rsa-sha256; c=relaxed/relaxed; d=gmail.com; s=20120113; h=mime-version:date:message-id:subject:from:to:content-type; bh=Q1YZ+t7iVg0WGVm9gpyG3yplQIC2Hy8K32YxfSjqieE=; b=Nw++Tlfp1StLFOvlLvuvCS8jluy9UcbPBtlvnssbnFqNEpR7YE-hcTw4SF+sNjQ4fZE MTkq0ywGxfgOUgUsc5v1KO3QYceKb5gbSJn3Q8ttxS0-TA3RhX64qUFezmsCvxm/7DXxb fczOpRvDzygRtazfk6VMWH9HK+t4jja6PSwFaHLle0YT-/pAhdoiLYJOxpdFgqNN08Qn GqzEU48FC5RuaBCt2tY+I7ek12sdQ3NNi6Pu/WfRqEsPgK3-FiDDFCR0HjY+QuoNPS2St0 Q/USE7SHr6g9/WAB3OTgd6b+kMSoedXgeb1MQU44sx-KqbJLKbBT4uaqyLPjE2E+td3th YScg==
MIME-Version: 1.0
X-Received: by 10.180.90.106 with SMTP id bv10mr27150189wib.4.1428004358345; Thu, 02 Apr 2015 12:52:38 -0700 (PDT)
Received: by 10.27.92.204 with HTTP; Thu, 2 Apr 2015 12:52:38 -0700 (PDT)
Date: Thu, 2 Apr 2015 21:52:38 +0200
Message-ID: <CAOkV4HnhJ4cn53rMysh9QGVaqmBd3RU-v=3L=8w1eMB+m27mLqw@mail.gmail.com>
Subject: fattura per lavori Marzo
From: Giovanni Cino <giovanni.cino@gmail.com>
X-ASG-Orig-Subj: fattura per lavori Marzo
To: "amministrazione@hackingteam.it" <amministrazione@hackingteam.it>
Content-Type: multipart/mixed; boundary="f46d043c81e88486140512c32c54"
X-Barracuda-Connect: mail-wi0-f178.google.com[209.85.212.178]
X-Barracuda-Start-Time: 1428004358
X-Barracuda-URL: http://192.168.100.25:8000/cgi-mod/mark.cgi
X-Virus-Scanned: by bsmtpd at hackingteam.com
X-Barracuda-BRTS-Status: 1
X-Barracuda-Spam-Score: 0.00
X-Barracuda-Spam-Status: No, SCORE=0.00 using global scores of TAG_LEVEL=3.5 QUARANTINE_LEVEL=1000.0 KILL_LEVEL=8.0 tests=HTML_MESSAGE
X-Barracuda-Spam-Report: Code version 3.2, rules version 3.2.3.17497
Rule breakdown below
pts rule name description
---- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------
0.00 HTML_MESSAGE BODY: HTML included in message
Return-Path: giovanni.cino@gmail.com
X-MS-Exchange-Organization-AuthSource: EXCHANGE.hackingteam.local
X-MS-Exchange-Organization-AuthAs: Internal
X-MS-Exchange-Organization-AuthMechanism: 10

Our **Database**

From	To	Subject	Date	Sender IP address
Giovanni Cino <giovanni.cino@gmail.com>	amministrazione@hackingteam.it	fattura per lavori Marzo	Thu, 2 Apr 2015 21:52:38 +0200	209.85.212.178

https://labs.rs/en/metadata/

Building a network from client side

- Email networks are generally weighted directed networks, since people can send more than one message

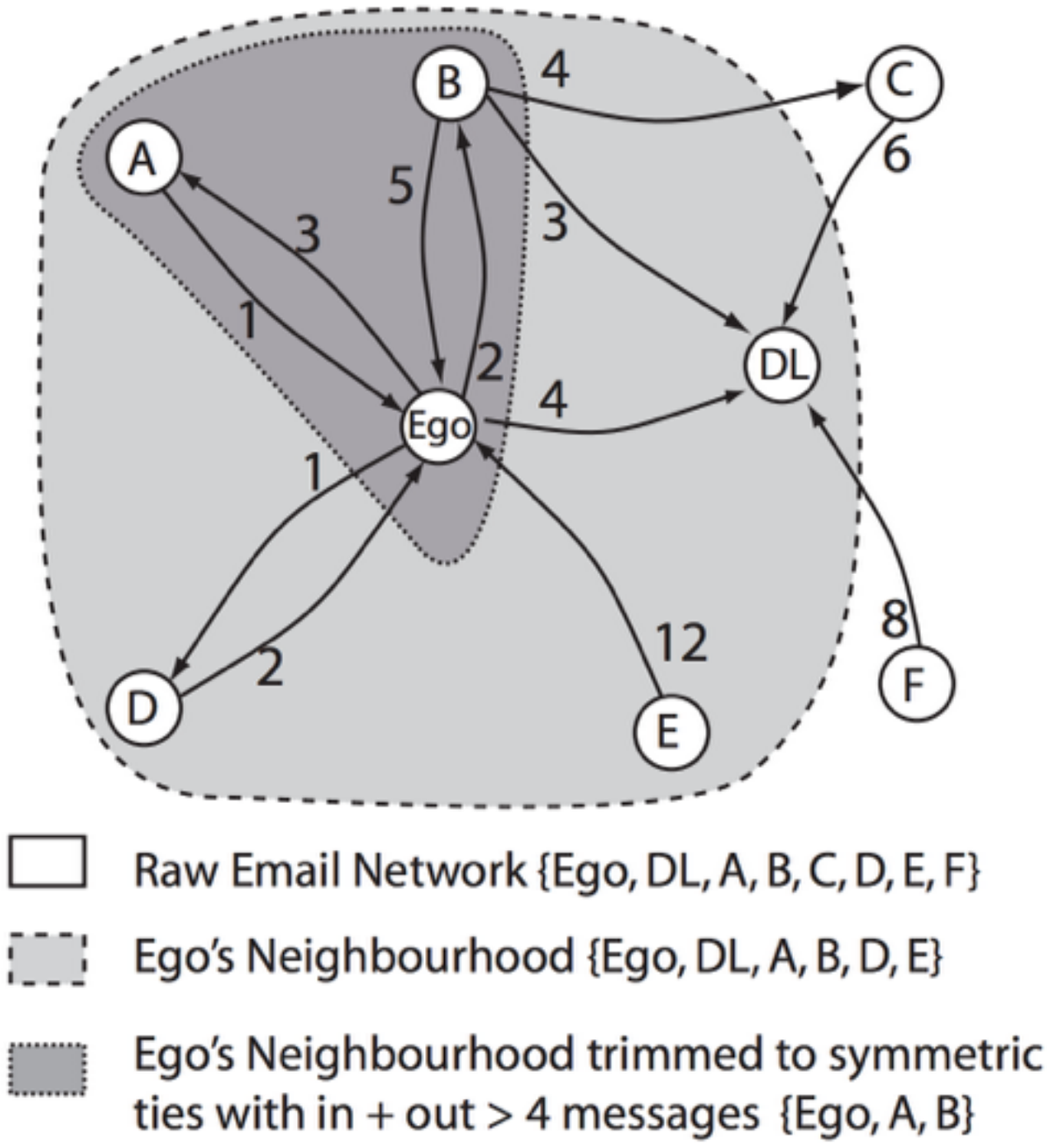
- Edges (arcs) go from the sender to each of the receivers

- Since messages are often sent to more than one person, and the recipients reply to everyone, there are often ties between the various email addresses in the mail store, and not just ties between ego (the owner of the mail store), and those people that send ego mail

Hogan, B. (2008) Analyzing Social Networks via the Internet. In: N.Fielding, R.Lee and G.Blank (eds) The Handbook of Online Research Methods. Thousand Oaks, CA: Sage, pp. 141-160.

Email thresholds

- The use of structural metrics allow to differentiate relevant correspondence from spam and mailing lists
- In this way, the network is trimmed down to specific messages
- Four possible nested threshold zones
  - Zone 1: All messages in a mail store
  - Zone 2: Ego's neighborhood (authors who have sent messages directly to ego, or received messages directly from ego)
  - Zone 3: Ego's symmetric neighborhood (there has to be a message from ego to alter and from alter to ego)
  - Zone 4: Ego's thresholded neighborhood (There has to be at least n messages from ego and (or) n messages from alter)



Graphic representation of these thresholds (with levels 3 and 4 collapsed into one zone)

Server side: Mailing lists

- A mailing list in an OSS project is a public forum
- Anyone can post messages to the list
- Posted messages are visible to all the mailing list subscribers
- Posters to mailing lists include developers, bug-reporters, contributors (who submit patches, but don't have commit privileges) and ordinary users

epiphany-list -- For developers and users of the Gnome web browser

About epiphany-list

To see the collection of prior postings to the list, visit the epiphany-list Archive

Using epiphany-list

To post a message to all the list members, send email to epiphany-list@gnome.

You can subscribe to the list, or change your existing subscription, in the sectio

Subscribing to epiphany-list

Subscribe to epiphany-list by filling out the following form. You will be sent e prevent others from gratuitously subscribing you. This is a hidden list, which n available only to the list administrator.

Your email address:	
Your name (optional):	

You may enter a privacy password below. This provides only mild security, but should prevent others from messing with your subscription. **Do not use a valuable password a** will occasionally be emailed back to you in cleartext.

Integrating Tab Bar with Address Bar

- *From*: "Mark E. Lee" <mark markelee com>
- *To*: epiphany-list gnome org
- *Subject*: Integrating Tab Bar with Address Bar
- *Date*: Mon, 22 Apr 2013 23:29:18 -0400

```
The tab bar group wastes screenspace. Is it possible to move the tab bar
group to the side of the Address Bar (like how IE9 does it)?

Thanks,
Mark
--
Mark E. Lee <mark markelee com>
```

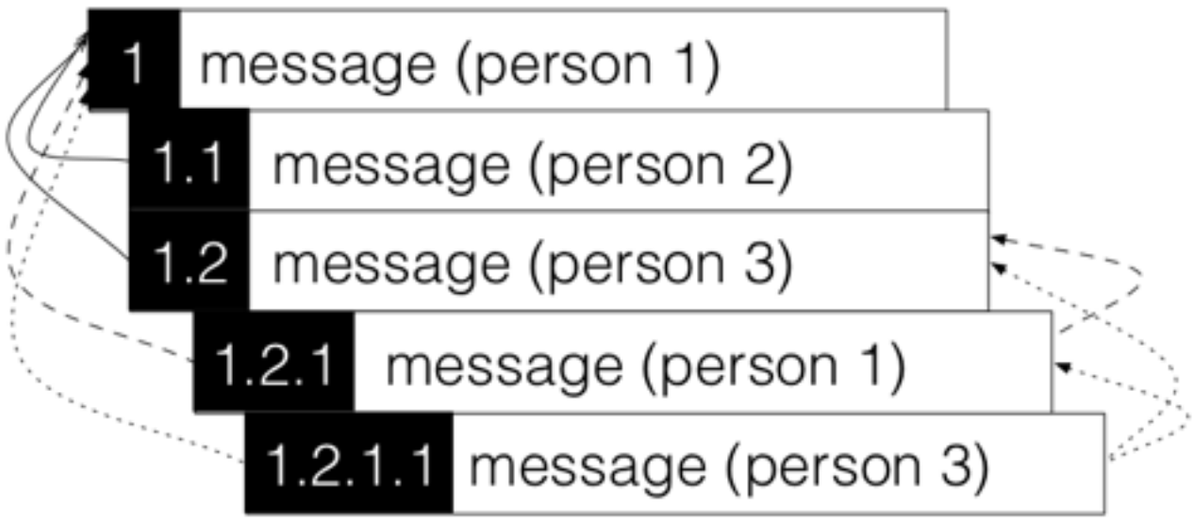
Attachment: signature.asc

Description: This is a digitally signed message part

[Date Prev][Date Next] [Thread Prev][Thread Next] [Thread Index] [Date Index] [Author Index]

Defining networks

- A response b to a message a is an indication that the sender of b, (s_b) found that the sender of a,(s_a) had something interesting to say
- The response from s_b indicates that the original message a represented information flowing from s_a to s_b
- It is also an indication of status, i.e., s_b indicates that s/he found s_a's email worth reading, and worthy of response
- Another example



	1	2	3
1	1	0	1
2	1	0	0
3	3	0	1

Unmasking aliases

- Problem:
  - People use more than one email address in projects, for example ian.holsman@cnet.com, ianh@holsman.net, and ianh@apache.org

- Approach
  - Automatically crawl messages and extract all such headers to produce a list of < name,email > identifiers (IDs)
  - Executing of a clustering algorithm that measures the similarity between every pair of IDs
  - Clustering is carried out if either the names are similar, or if the emails are similar, or if the names and the emails are similar
  - IDs that are sufficiently similar are placed into the same cluster
  - Each cluster is manually post- processed

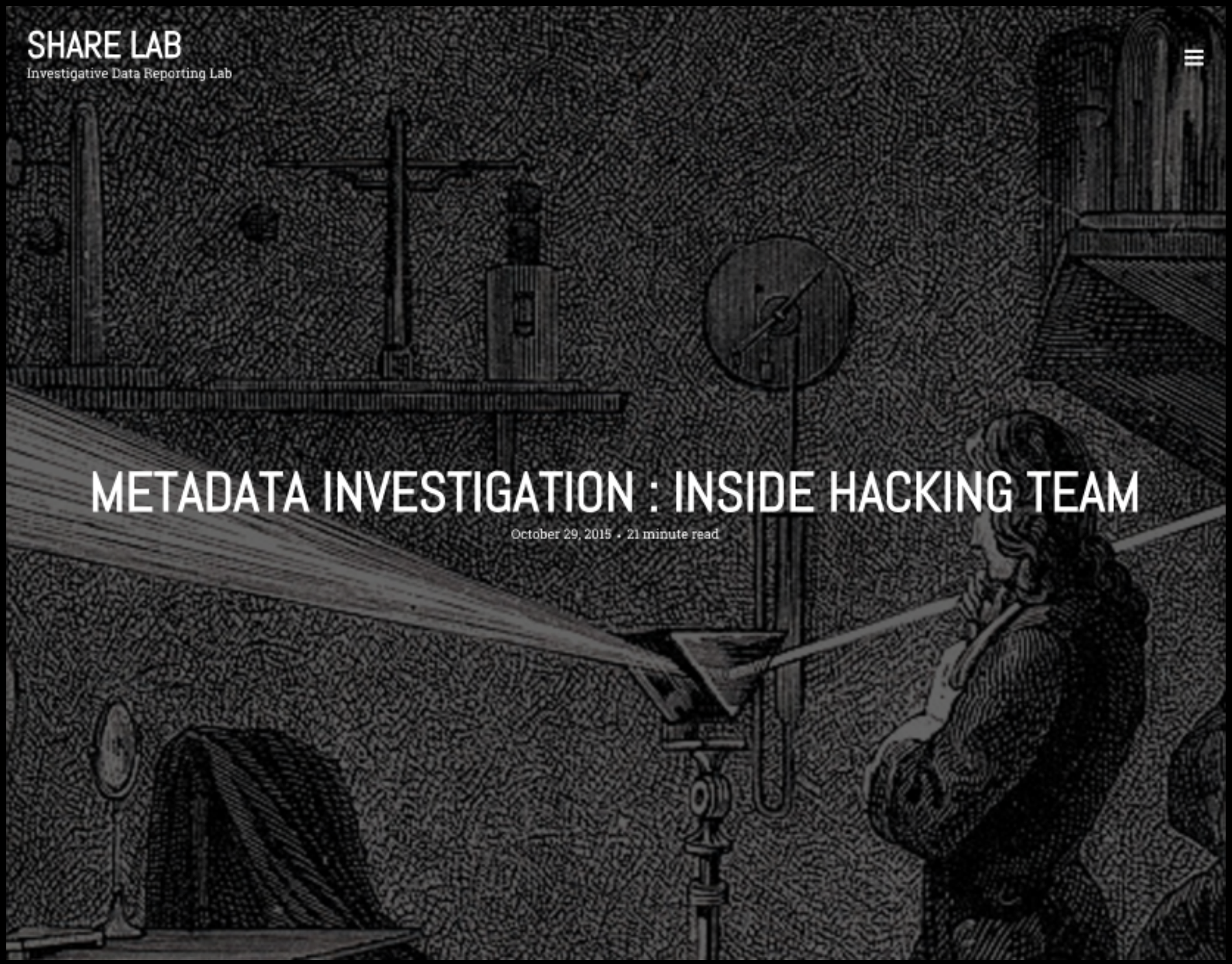
- Tipp: Set the cluster similarity threshold quite low: it is much easier during a manual step to split clusters than to unify two disparate clusters from a very large set

Bird, C., Gourley, A., Devanbu, P., Gertz, M., & Swaminathan, A. 2006, 'Mining email social networks', 137--143.

Unmasking aliases - Clustering Algorithm

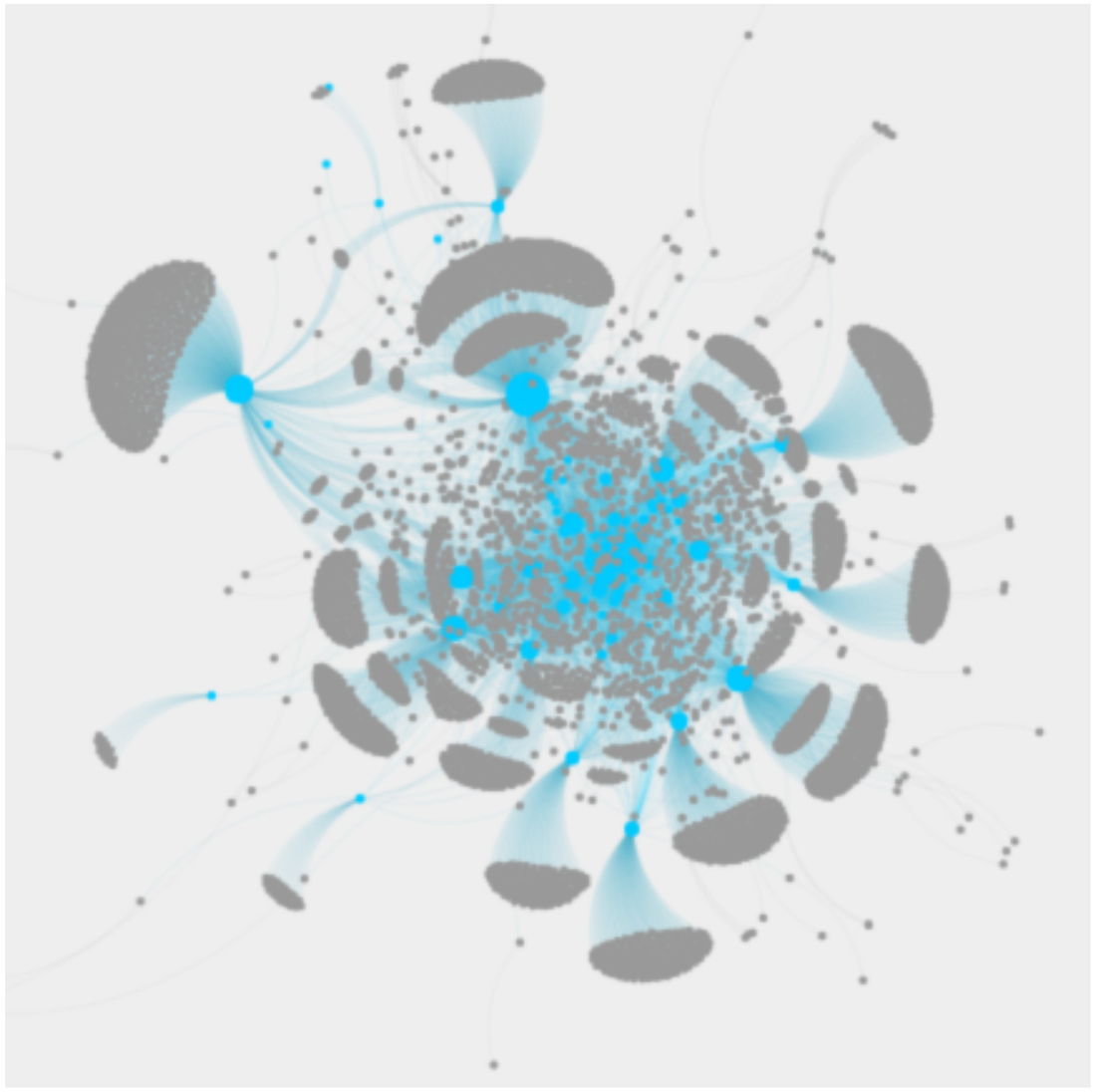
- Input: list of IDs (a < name,email > tuple)
- Step 1: Create pairwise similarity measures for every pair of IDs
- Step 2: Place two IDs into the same cluster if the similarity measure exceeds a threshold
- Computation of the similarity measure contains:
  - Normalize name (e.g., remove all punctuation, suffixes (“jr”), turn all whitespace into a single space)
  - Name similarity: use a scoring algorithm based on the Levenshtein edit distance (for full names, and the first and last names separately)
  - Names-email similarity: Two IDs are also scored highly similar if the emails and names match. For example, Erin Bird matches erinb and ebird
  - Email similarity: If the Levenshtein edit distance between two email address bases (not including the domain) is small
  - Cumulative ID similarity: The similarity between two IDs is the maximum of the 3 mentioned above

Bird, C., Gourley, A., Devanbu, P., Gertz, M., & Swaminathan, A. 2006, 'Mining email social networks', 137--143.

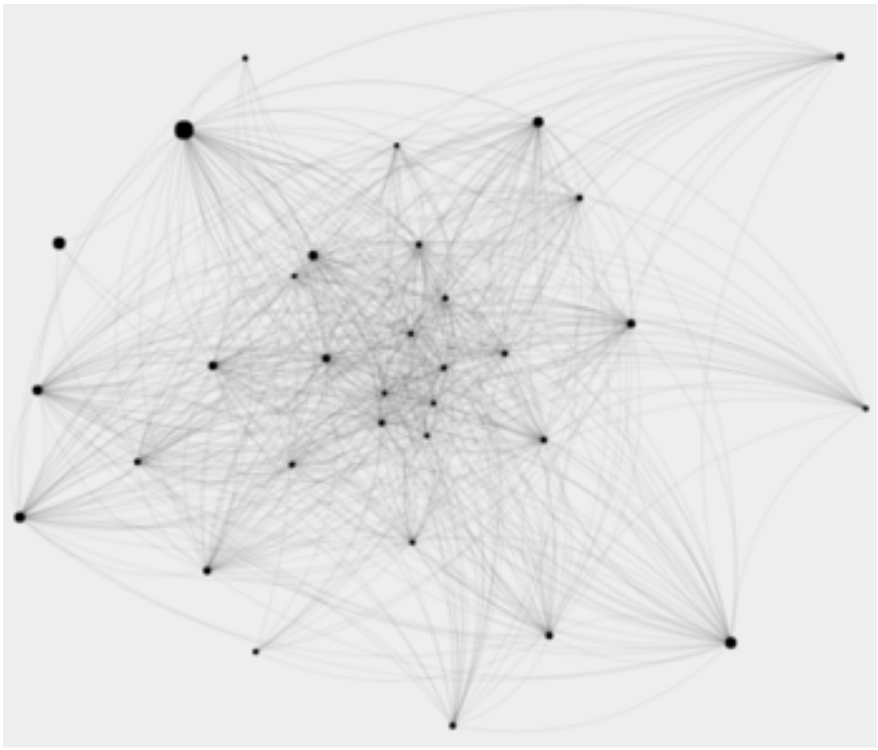


METADATA INVESTIGATION : INSIDE HACKING TEAM

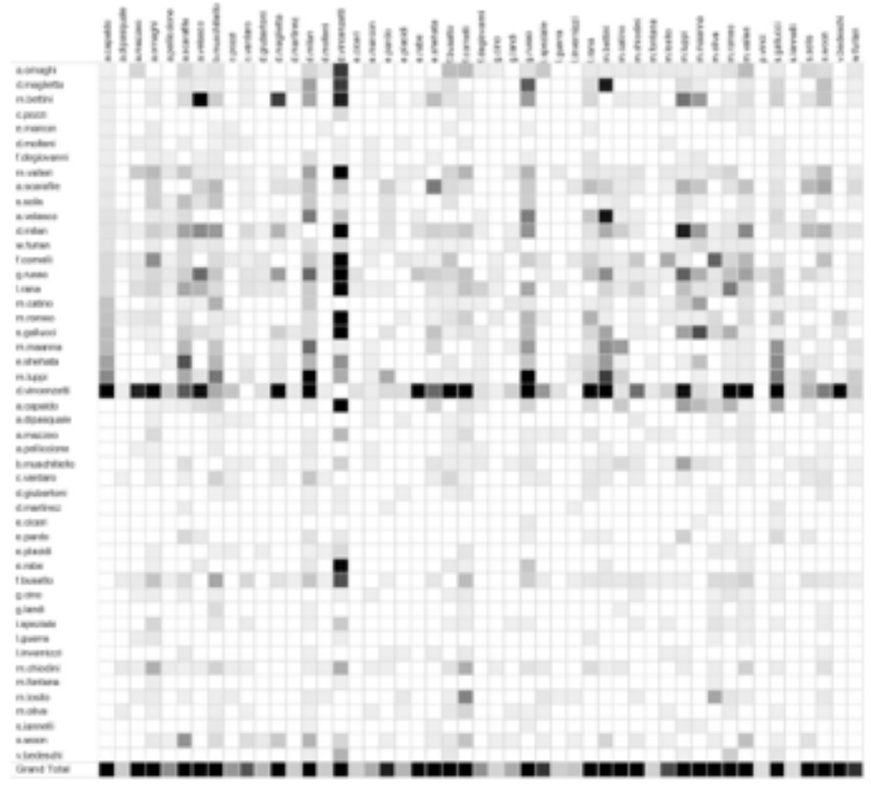
October 29, 2015 · 21 minute read



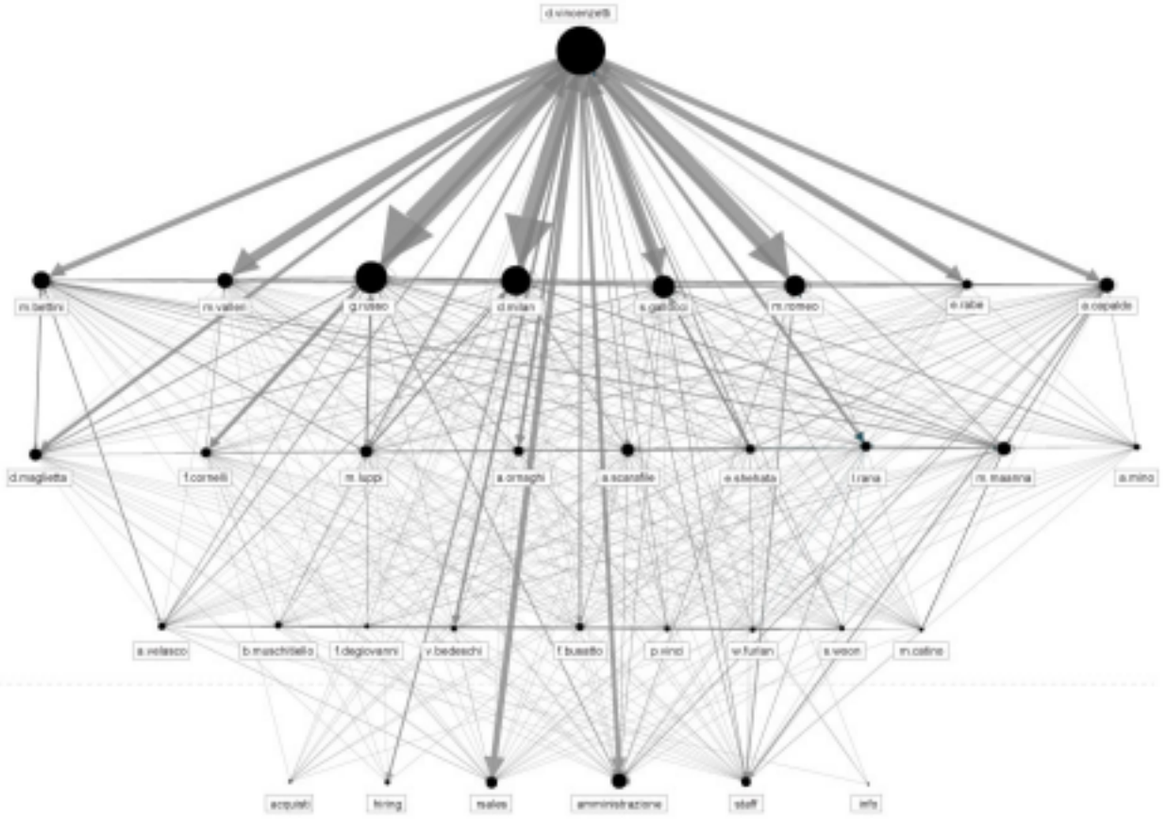
SOCIAL NETWORK ANALYSIS OF NODES WITH +100 EXCHANGED EMAILS



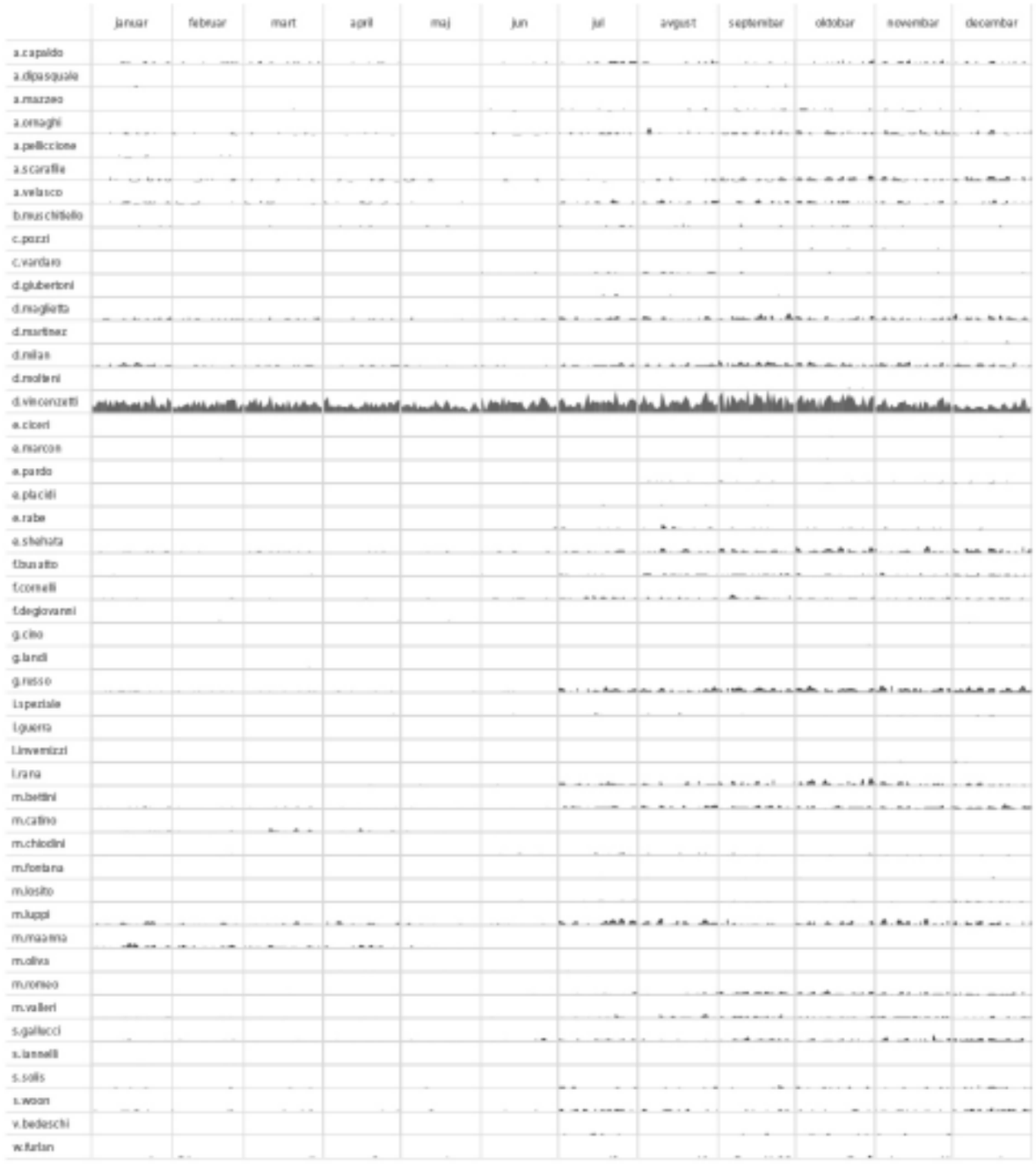
SOCIAL NETWORK ANALYSIS OF HACKING TEAM EMAIL DATABASE (PERIOD 2013-2015)



HEAT-MAP OF INTERNAL COMMUNICATION



POTENTIAL ORGANISATIONAL STRUCTURE BASED ON THE LEVEL AND DIRECTION OF COMMUNICATION



NUMBER OF SENT EMAILS PER HT EMPLOYEE IN TIME (2014)

Questions?

- Check out:
  - http://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf