

Social Data Mining

Dino Pedreschi

KDD LAB

Università di Pisa & ISTI-CNR

kdd.isti.cnr.it

Scuola Superiore di Studi Avanzati della Sapienza

Roma | 28 Feb 2012



*We live in a time with
unprecedented opportunities of
sensing, storing, analyzing*

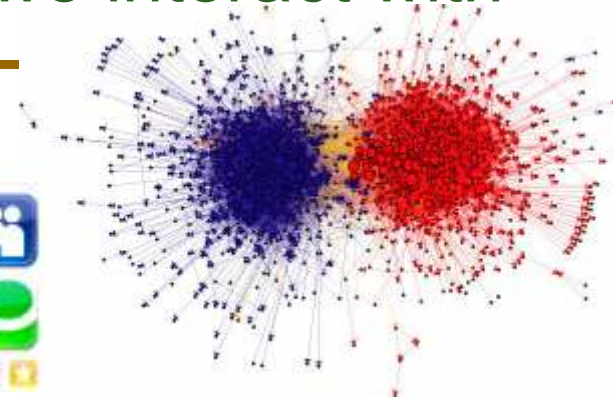
*(micro)-data, at mass level,
recording human activities
at extreme detail*



What we buy



Whom we interact with



What we search for



Where we go





98,000+
TWEETS



370,000+ MINUTES
VOICE CALLS ON

skype

12,000+
NEW ADS
POSTED ON
craigslist



13,000+ HOURS
MUSIC
STREAMING ON
PANDORA



1,600+
READS ON
Scribd.



1
NEW
DEFINITION
IS ADDED ON
URBAN



20,000+
NEW
POSTS ON
tumblr.



13,000+
iPhone
APPLICATIONS
DOWNLOADED

100+
Answers.com
40+
YAHOO! ANSWERS

QUESTIONS
ASKED ON THE
INTERNET...



600+
NEW
VIDEOS



70+
DOMAINS
REGISTERED

25+ HOURS
TOTAL
DURATION



60+
NEW
BLOGS

1,500+
BLOG
POSTS

168 MILLION
EMAILS
ARE SENT



694,445
SEARCH
QUERIES

Google

1,700+
Firefox
DOWNLOADS

695,000+
facebook
STATUS
UPDATES



79,364
WALL
POSTS



125+
PLUGIN
DOWNLOADS

510,040
COMMENTS



320+
NEW
twitter
ACCOUNTS



100+
NEW
Linked in
ACCOUNTS

1 associatedcontent
NEW
ARTICLE IS
PUBLISHED

THE
WORLD'S
LARGEST
COMMUNITY
CREATED CONTENT!!

6,600+
NEW
PICTURES ARE
UPLOADED ON
flickr



50+
WORDPRESS
DOWNLOADS

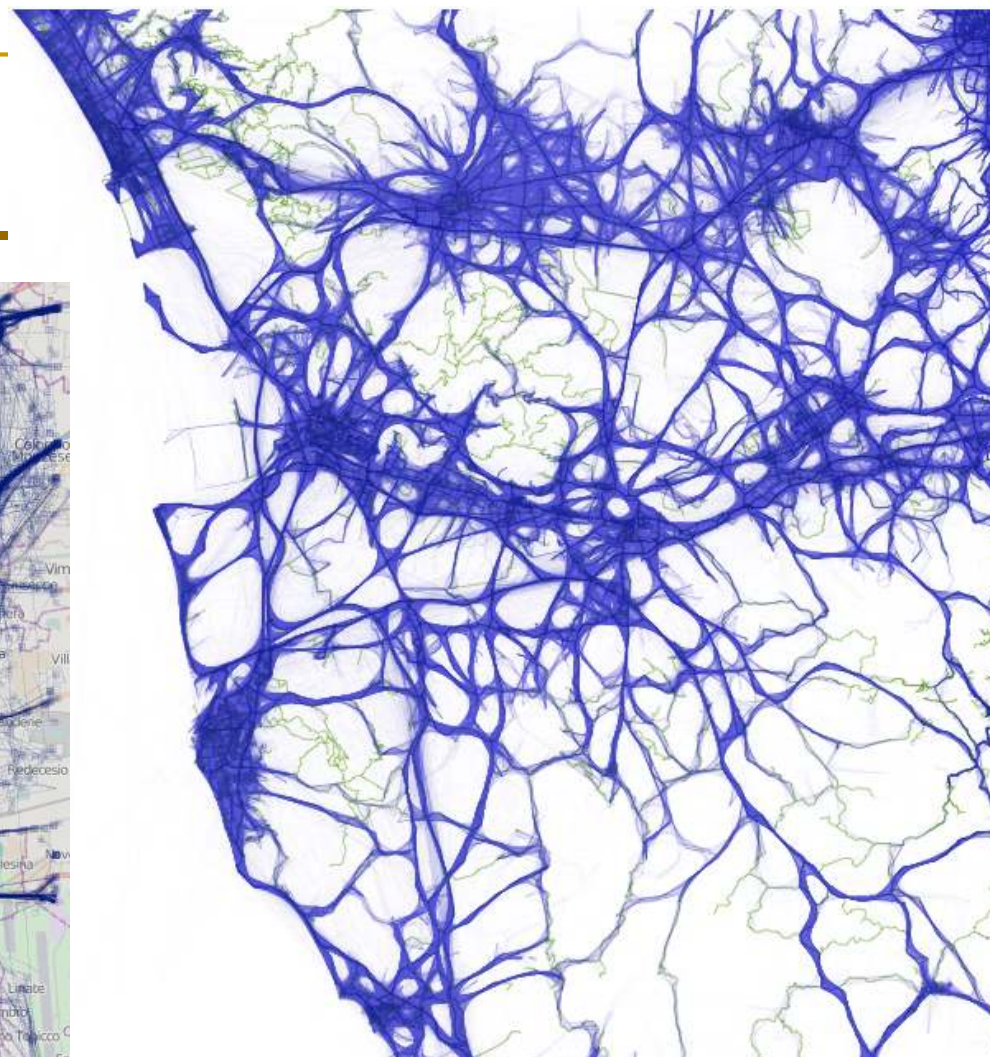
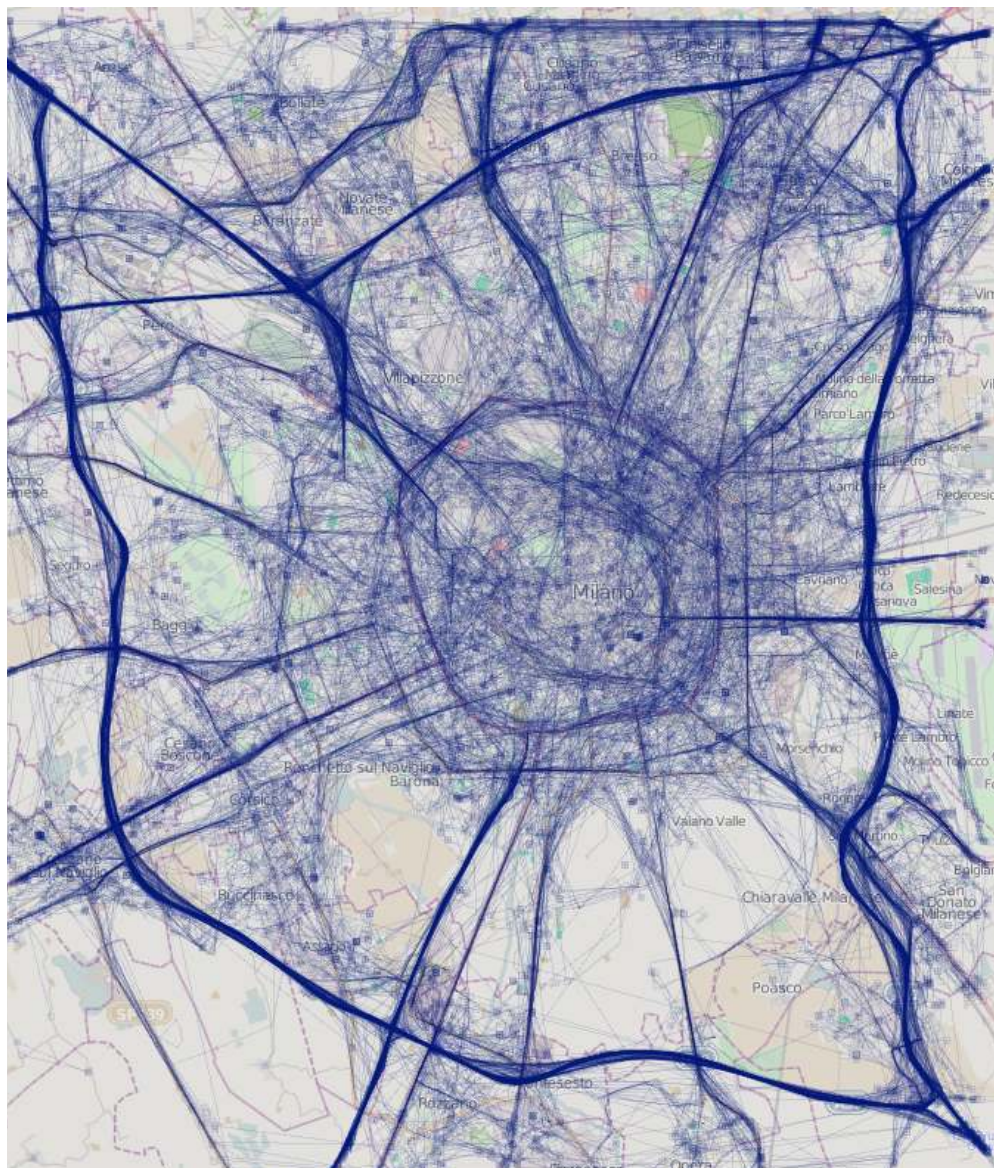
Paris

Fête de la Musique / mouvements des mobiles

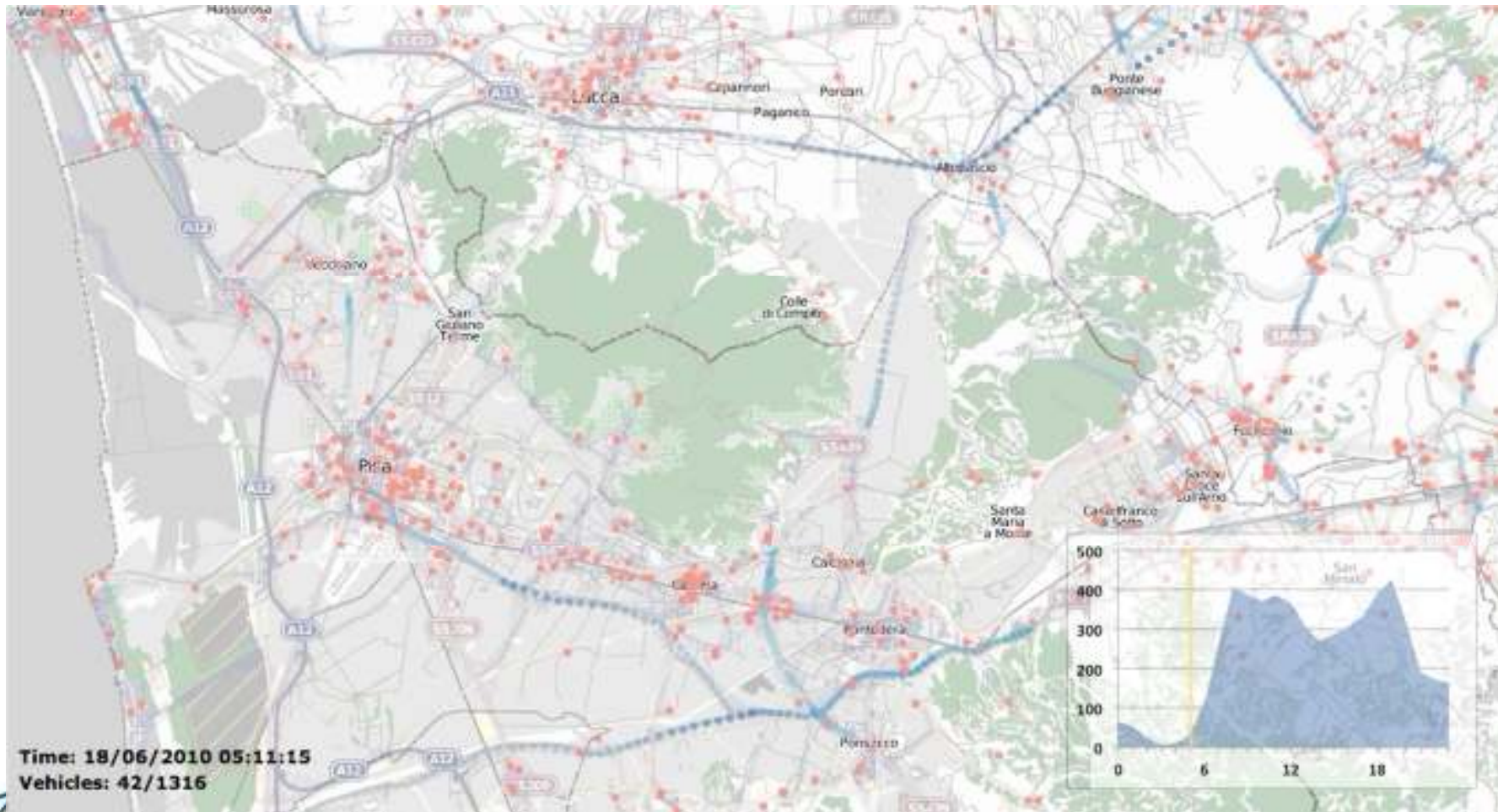
12:00 21/06/2008



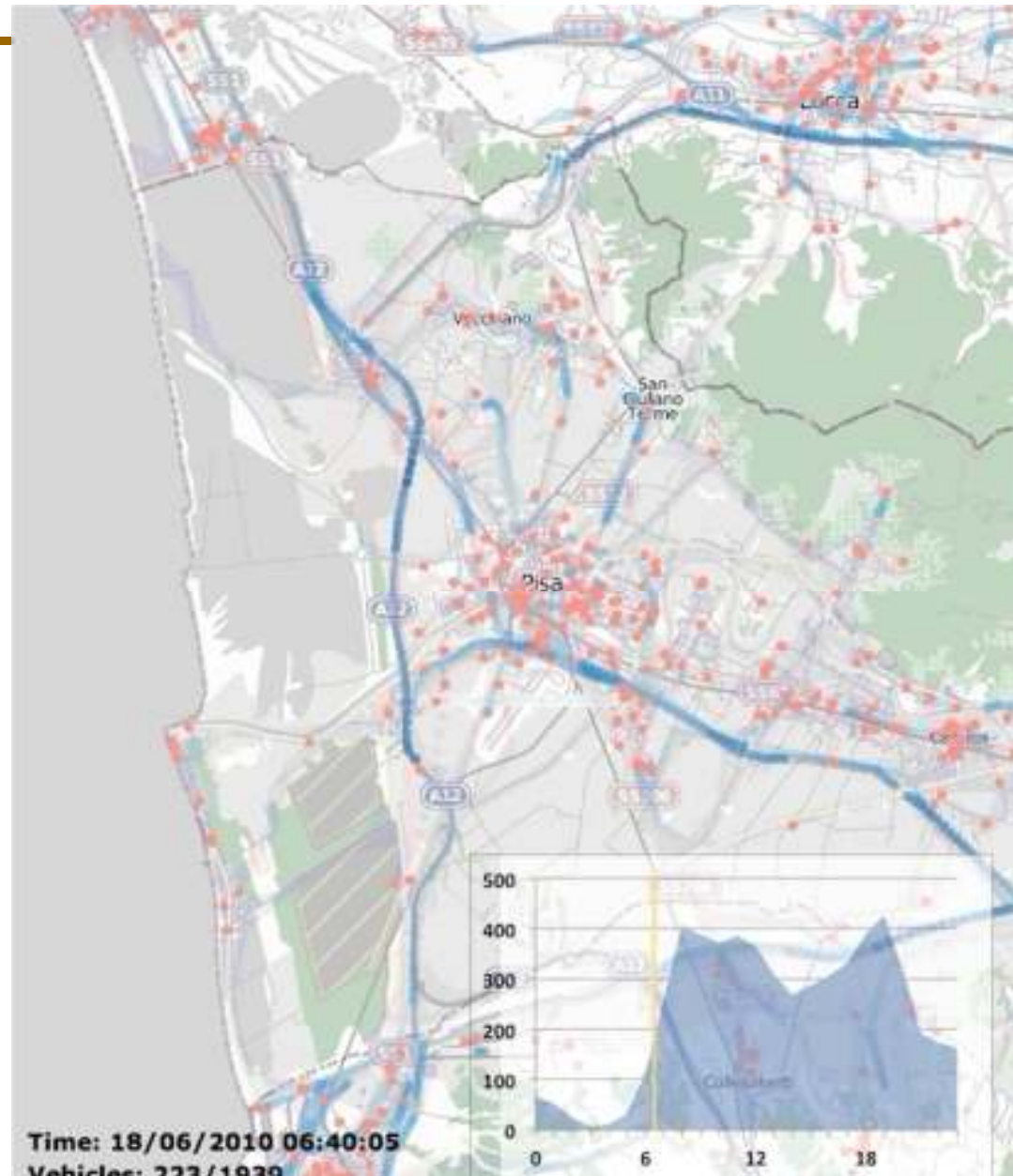
faberNovel



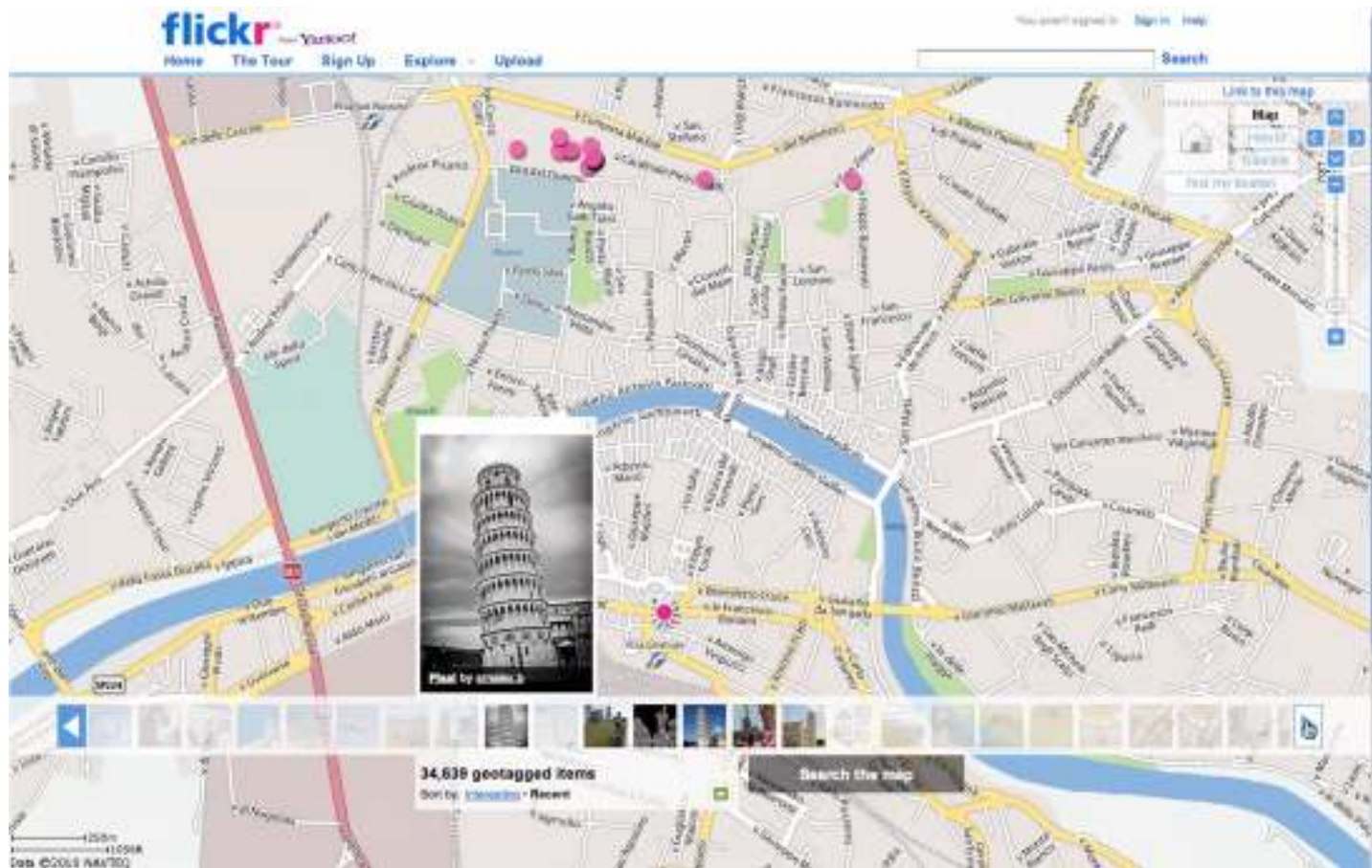
GPS: detailed movements within an area



GPS: movements within the town



Social Networks: goal of the movement



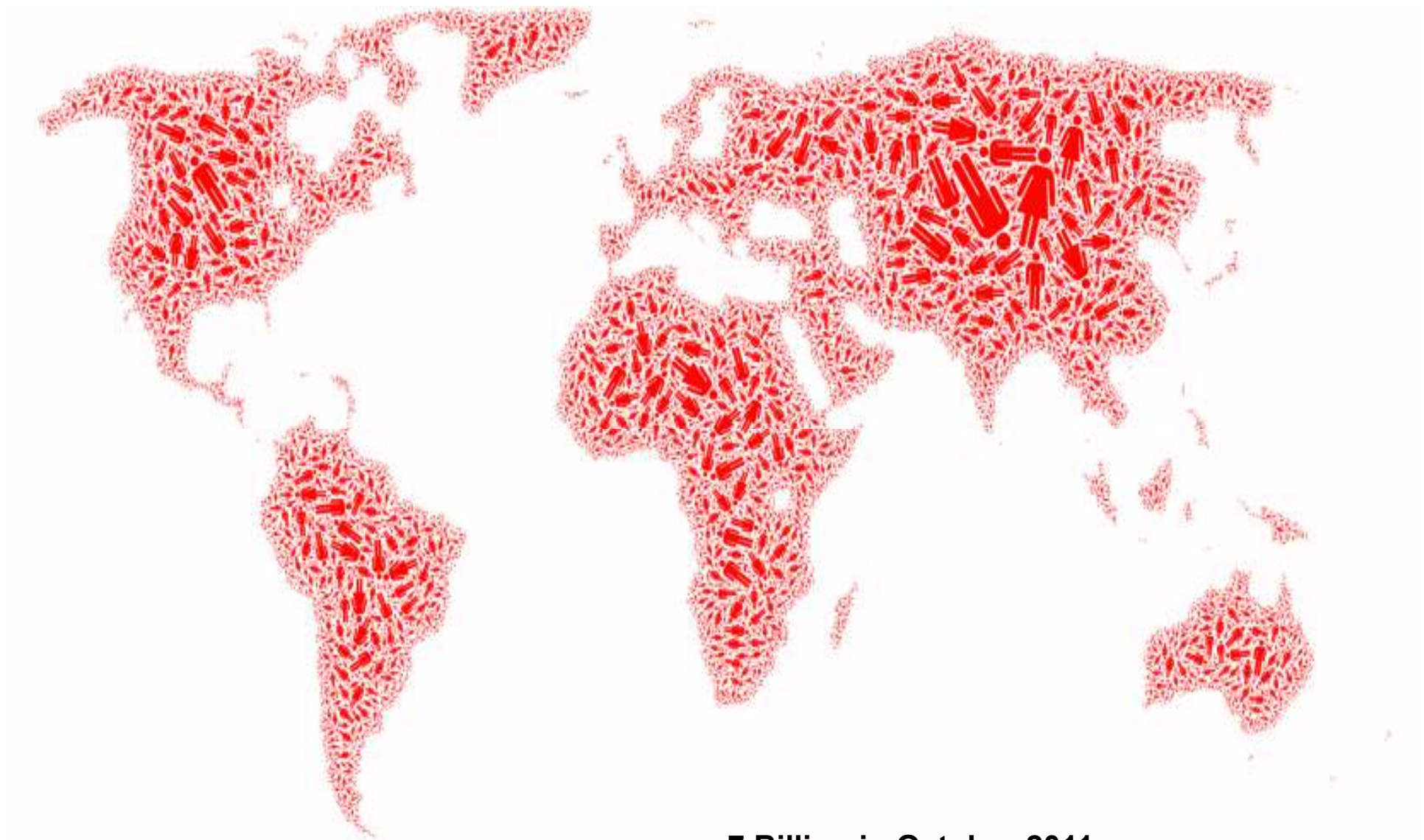


The social network behind Facebook

Big (human) data & social data mining

- a new social microscope
 - the basis of computational sociology
 - the basis of a data science (statistics 2.0)
- automated discovery of patterns, rules, profiles from human activity data
- The basis to understand and predict **complex phenomena** such as mobility, economic crises and trends, epidemics, opinion diffusion, sustainability





7 Billion in October 2011



A novel humanism

- The ICTs have revolutionized society
- Society is now revolutionizing the ICTs and scientific research as a whole
- Ultimately, the ICTs have brought humanity at the center of scientific investigation under a new light

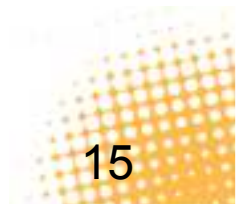


Social Data Mining

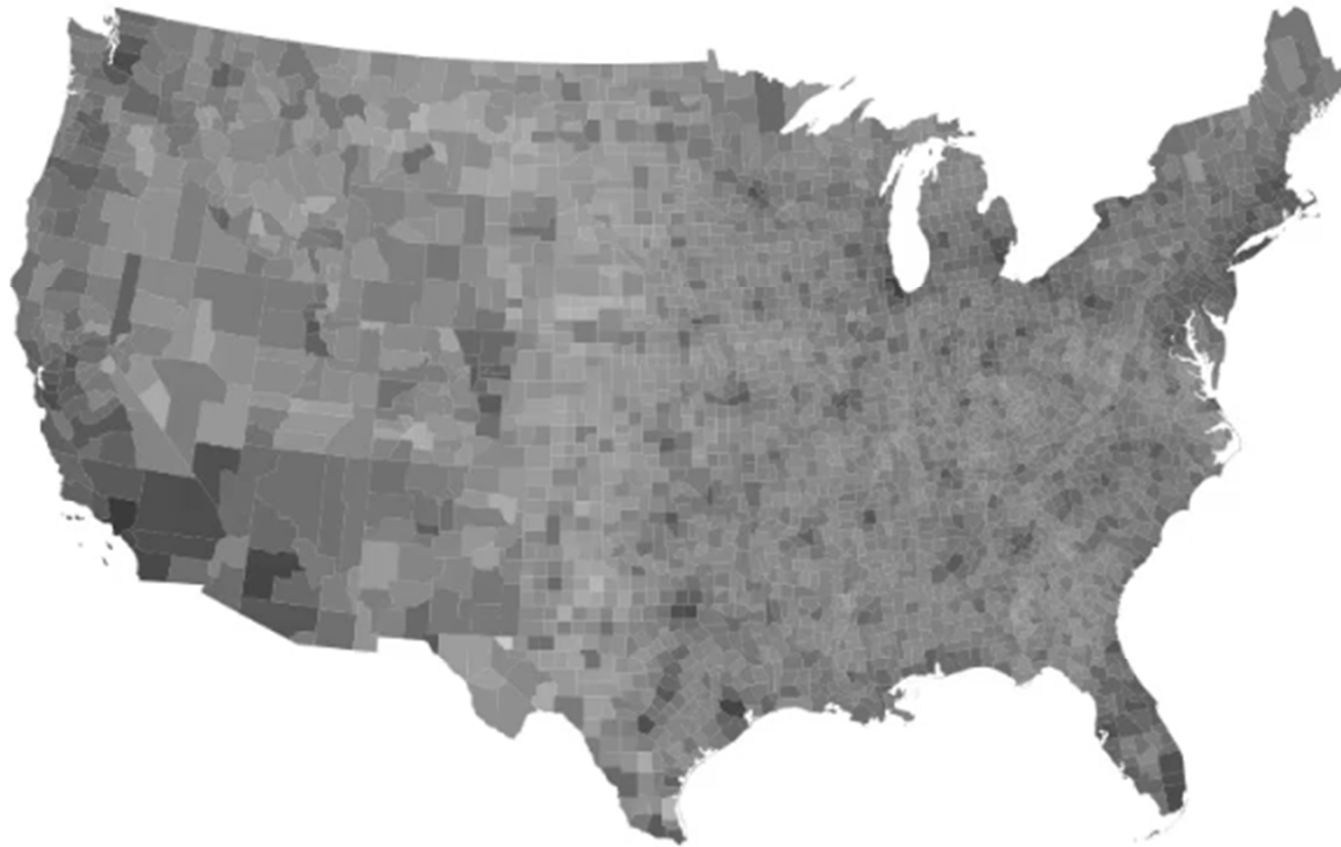


New questions to ask

-
- What are the geographical borders emerging from people's mobility and social behavior?



Mobility borders



Thiemann C, Theis F, Grady D, Brune R, Brockmann D, 2010
The Structure of Borders in a Small World. PLoS ONE 5(11)

Social borders

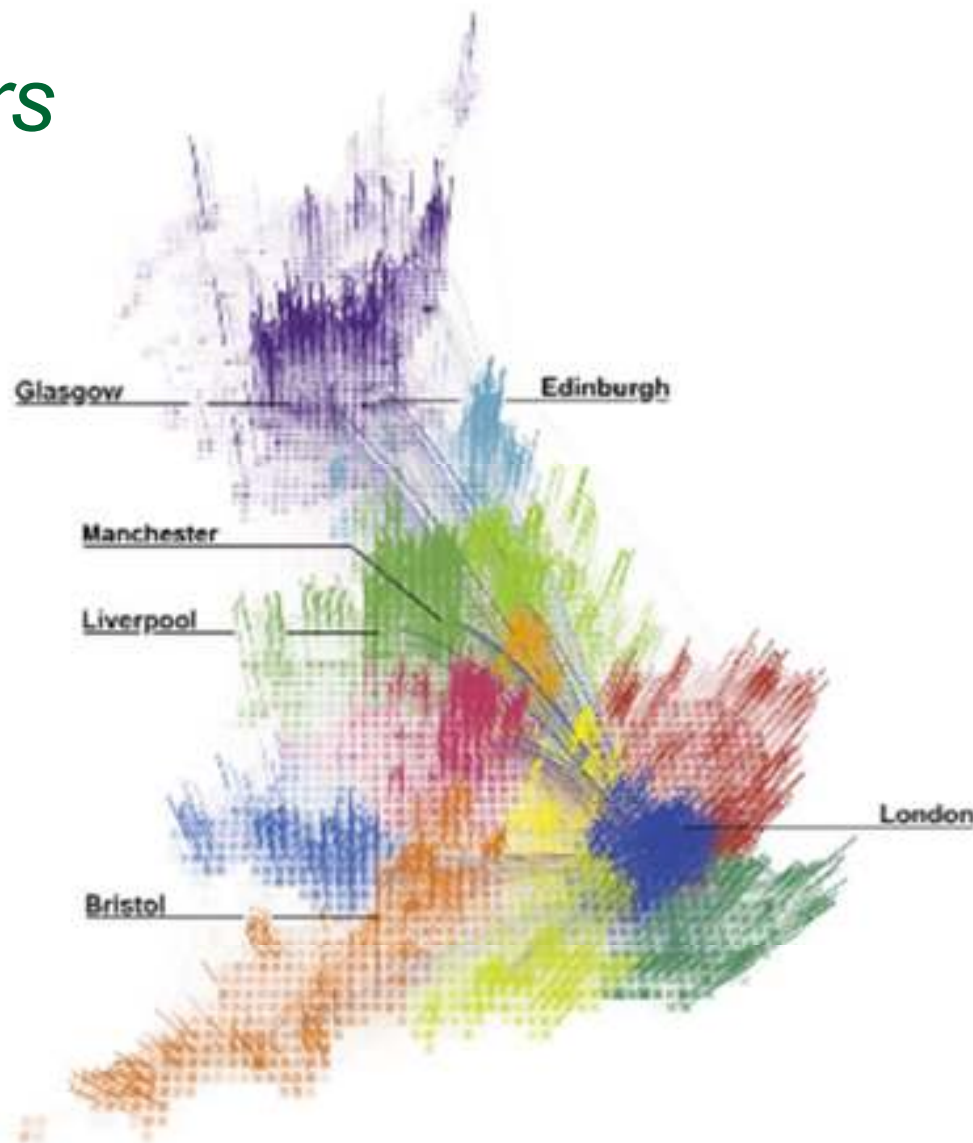


Figure 1. The geography of talk in Great Britain. This figure shows the strongest 80% of links, as measured by total 1 within Britain. The opacity of each link is proportional to the total call time between two areas and the different colours rep using network modularity optimisation analysis.
doi:10.1371/journal.pone.0014248.g001

Redrawing the map of Great Britain from a network of human interactions

Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R., Strogatz, S., Journal: PLoS ONE, 20

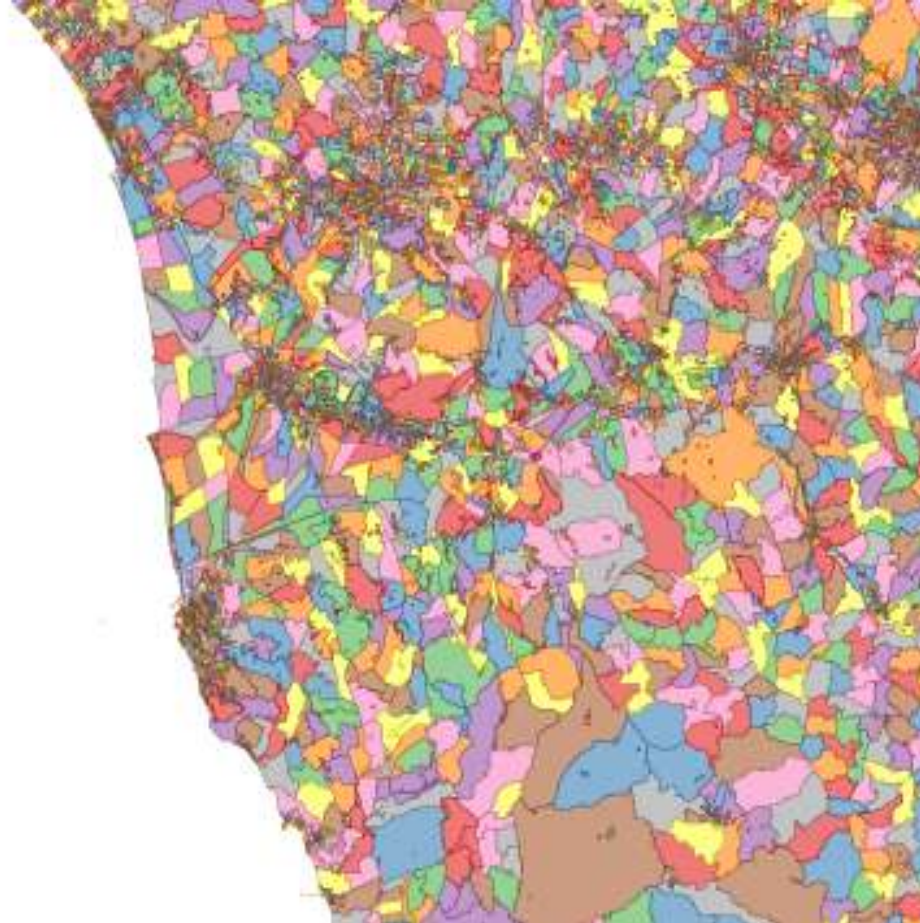


Networks as a mining tool

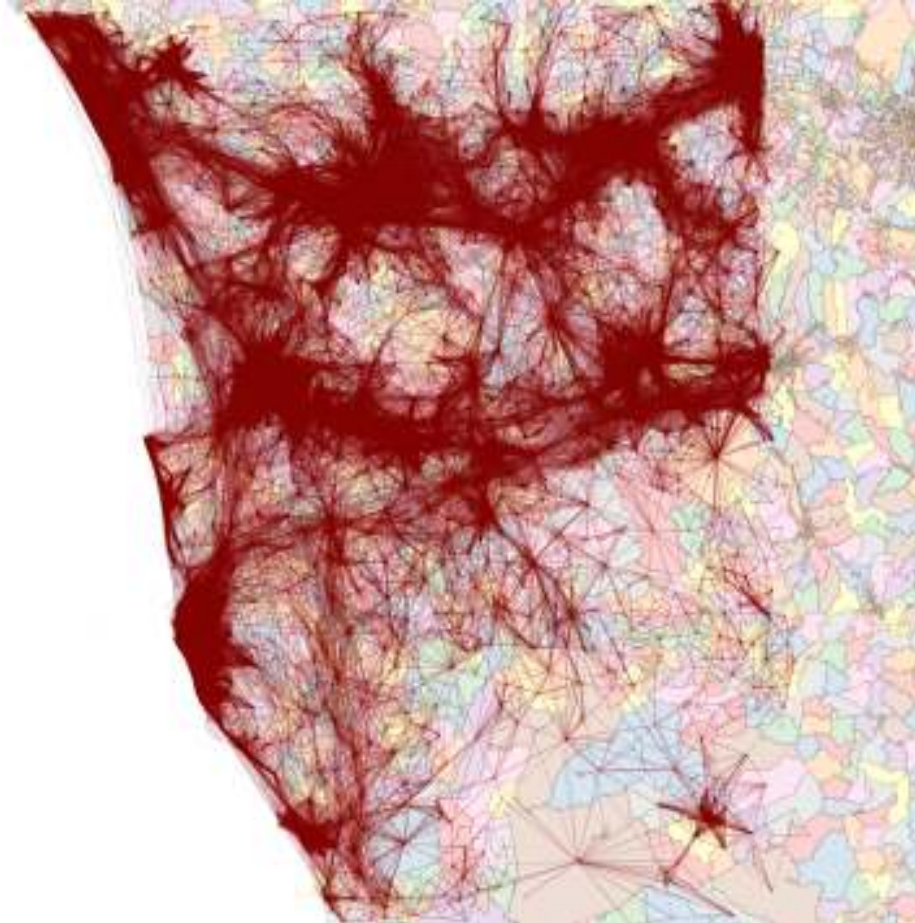


S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, F. Giannotti
Discovering the Geographical Borders of Human Mobility
KI - Künstliche Intelligenz, 2012 (to appear).

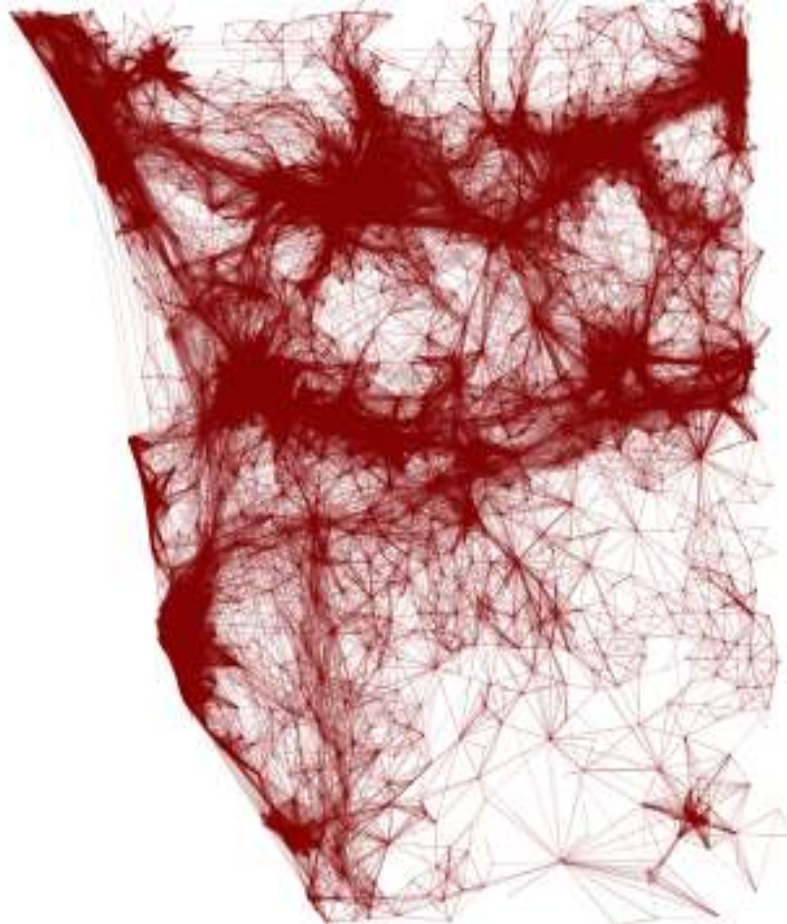
Step 1: spatial regions



Step 2: evaluate flows among regions



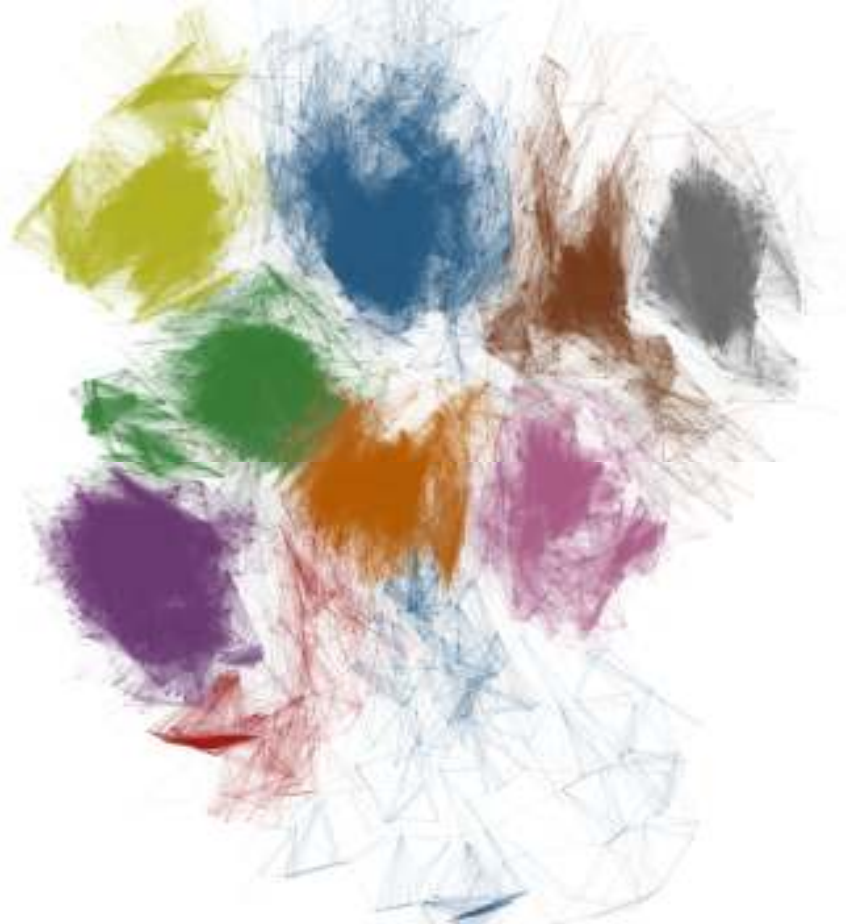
Step 3: forget geography



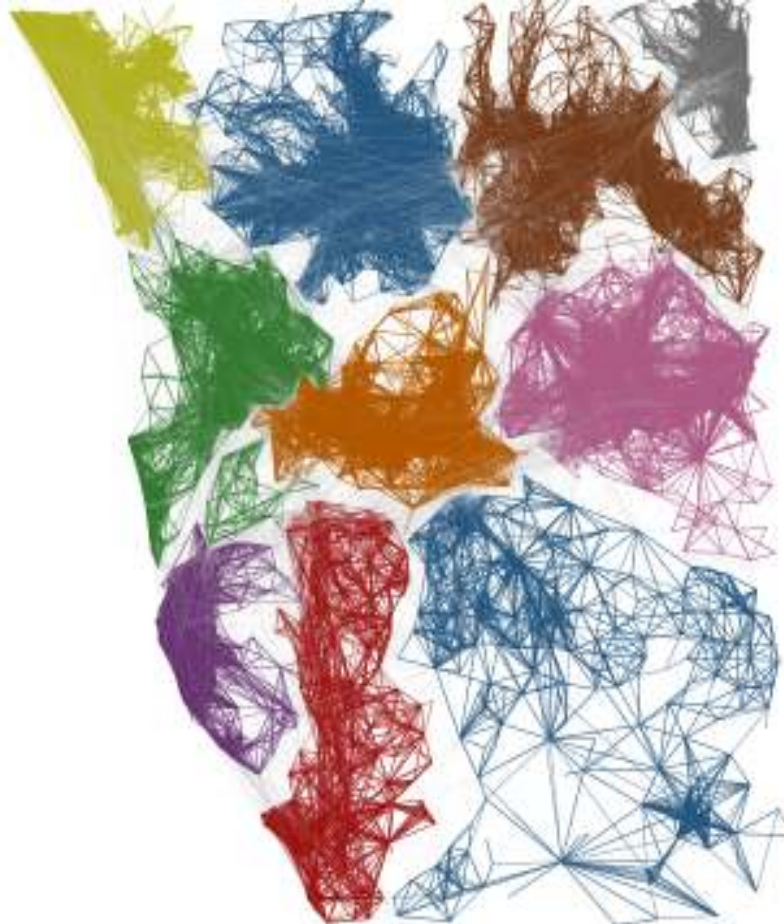
Step 4: perform community detection



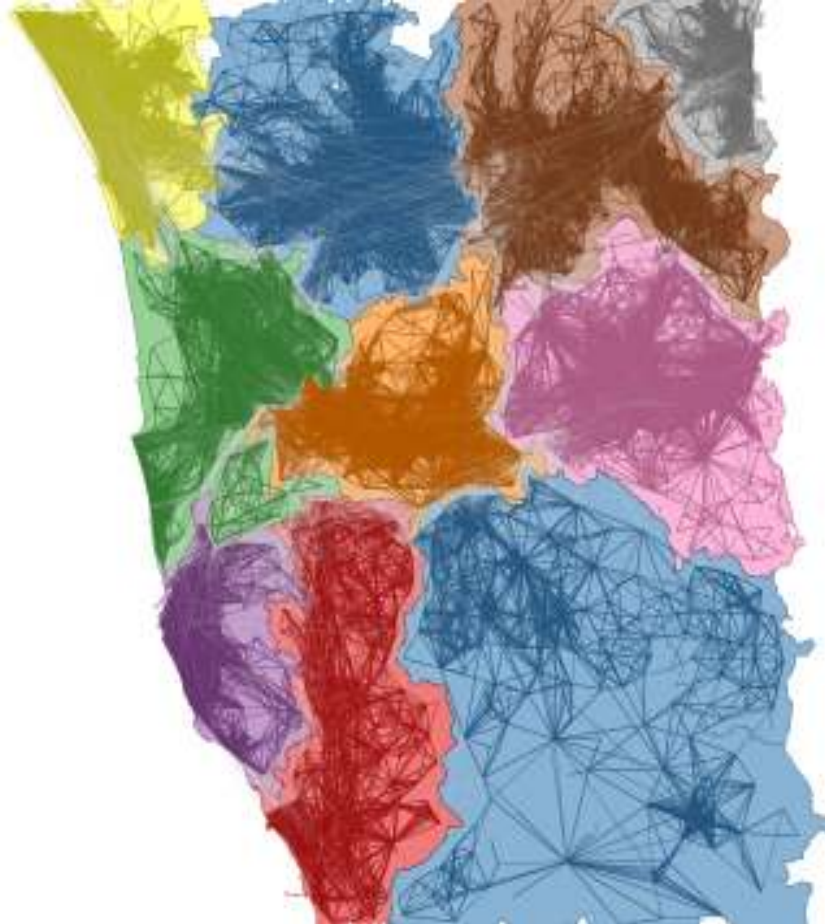
Step 4: perform community detection



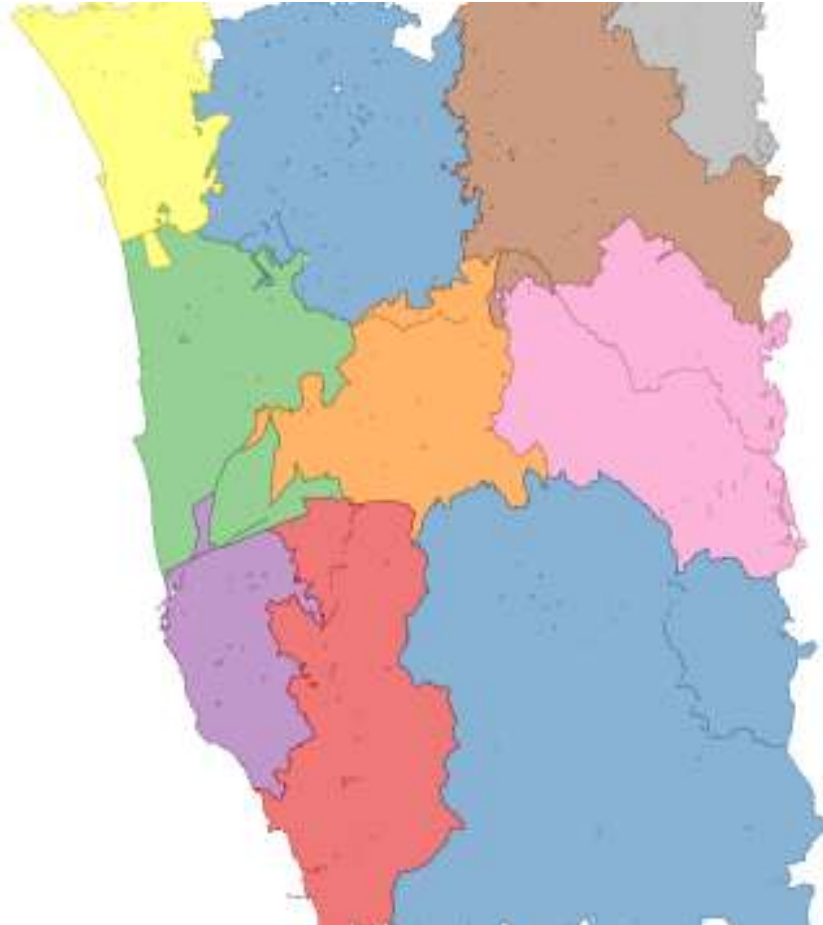
Step 5: map back to geography



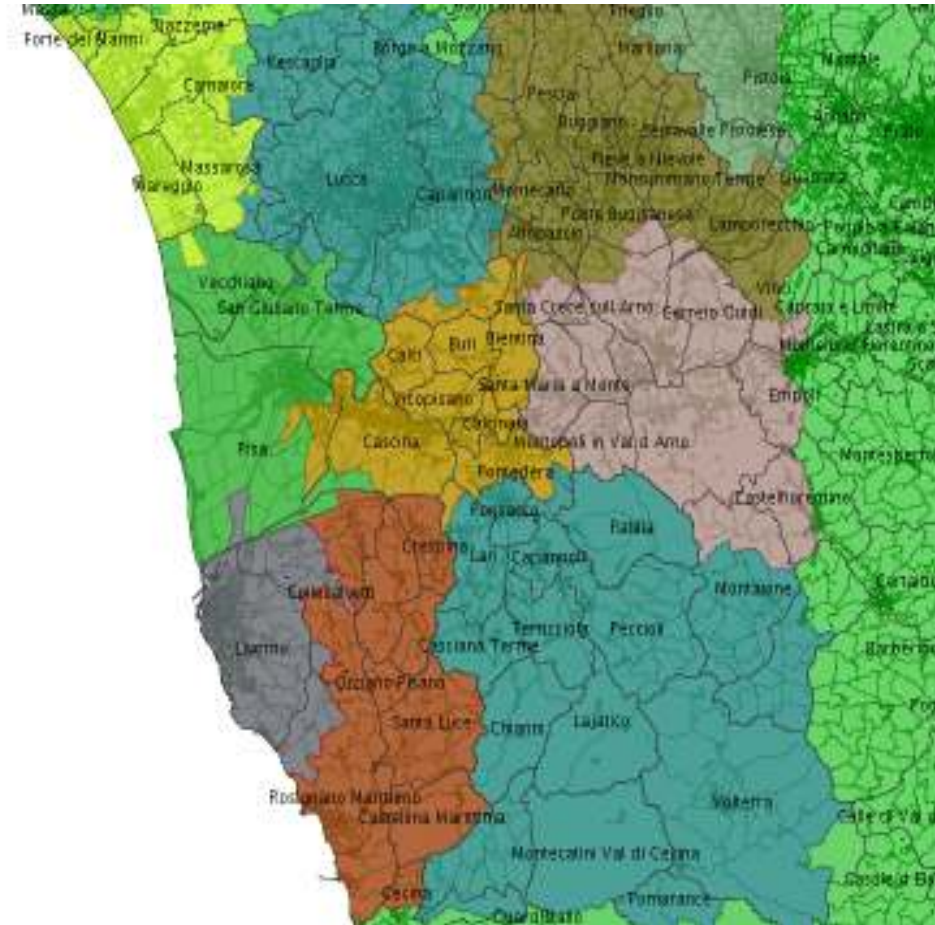
Step 6: draw borders



Final result



Final result: compare with municipality borders



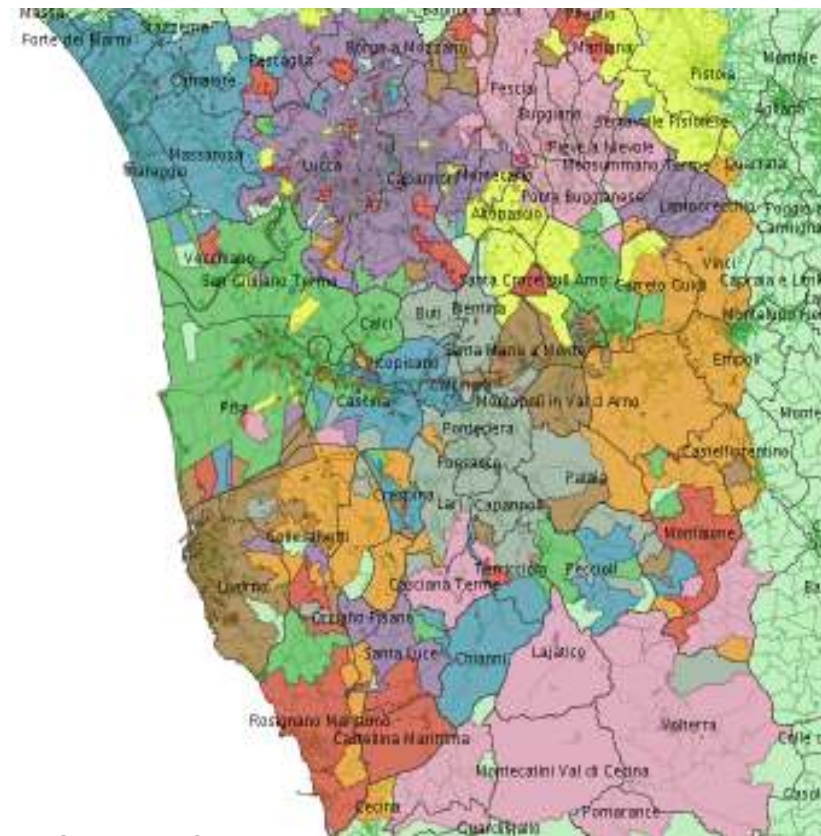
Borders in different time periods

Only weekdays movements



Similar to global clustering: strong influence of systematic movements

Only weekend movements

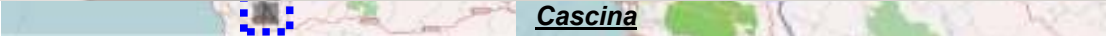


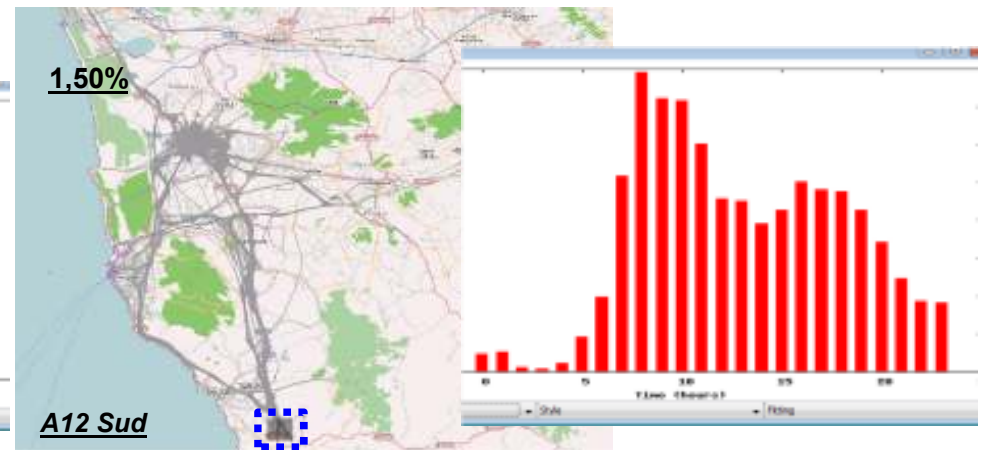
Strong fragmentation: the influence of systematic movements (home-work) is missing

-
- How people get into town?
What is the map of their
trips in space and time?

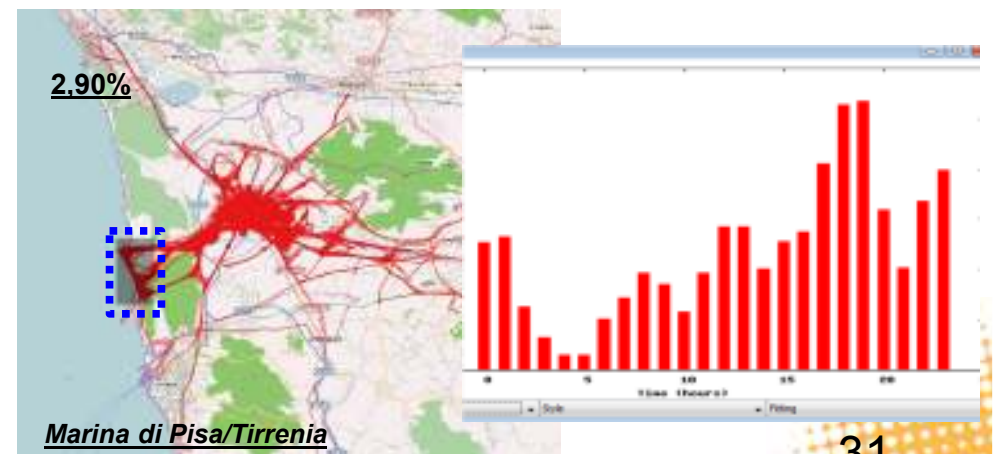
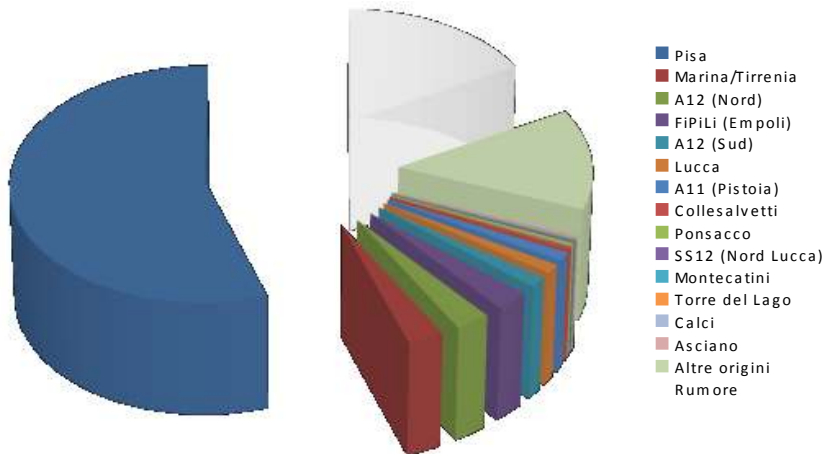


Understanding city access patterns

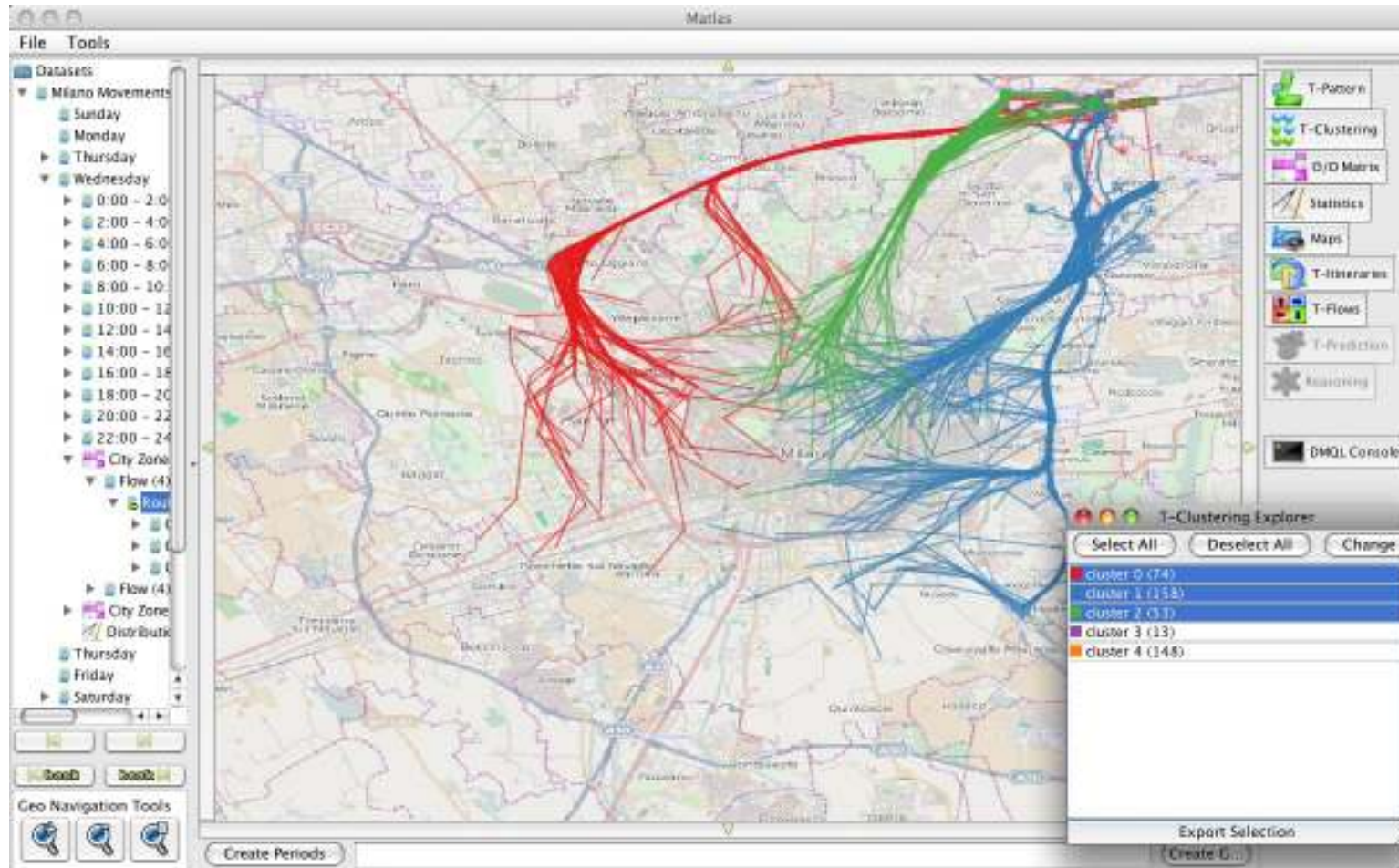




Distribuzione Origini



M-Atlas kdd.isti.cnr.it

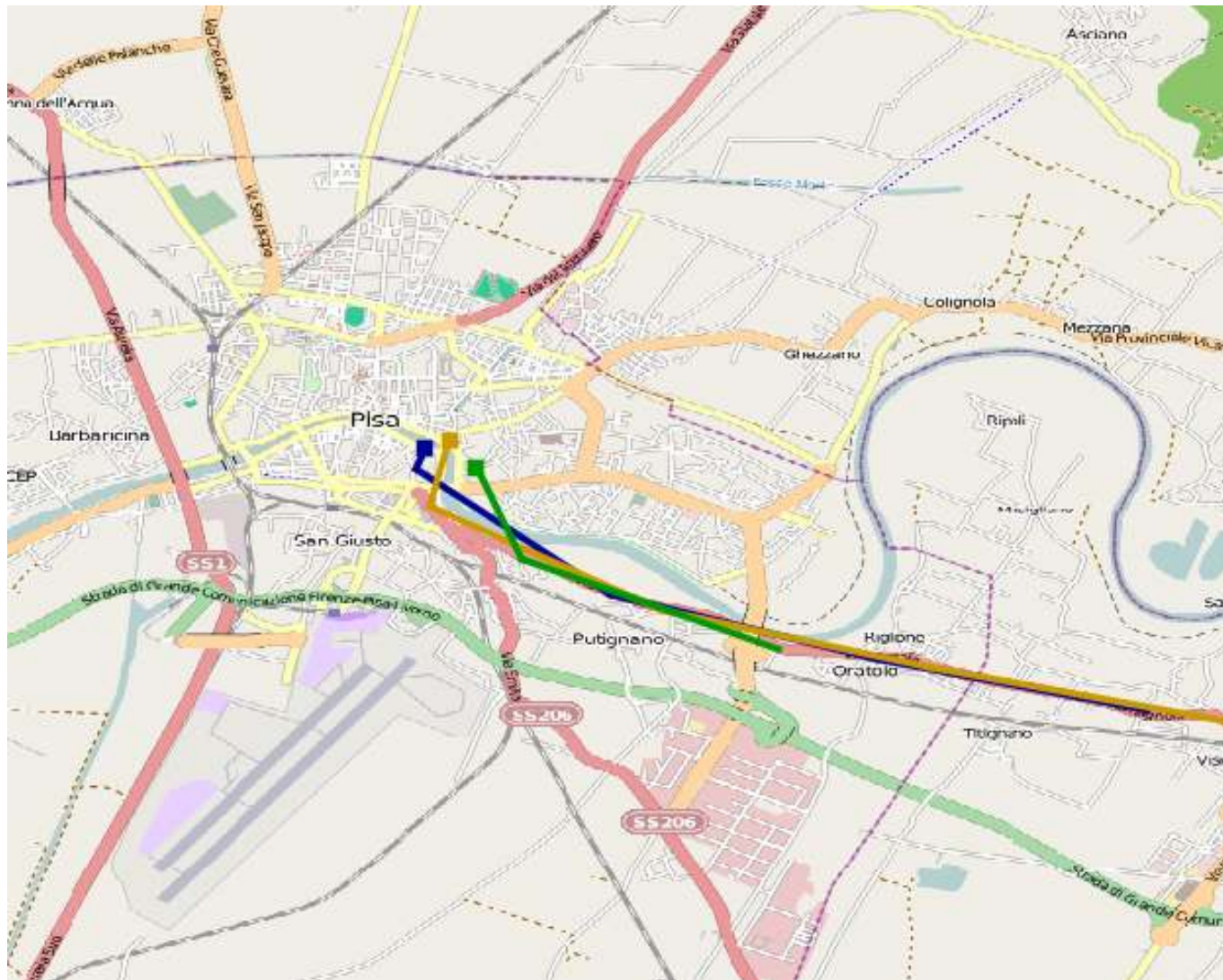


F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, R. Trasarti.
Unveiling the complexity of human mobility by querying and mining massive trajectory data.
VLDB J., 2011

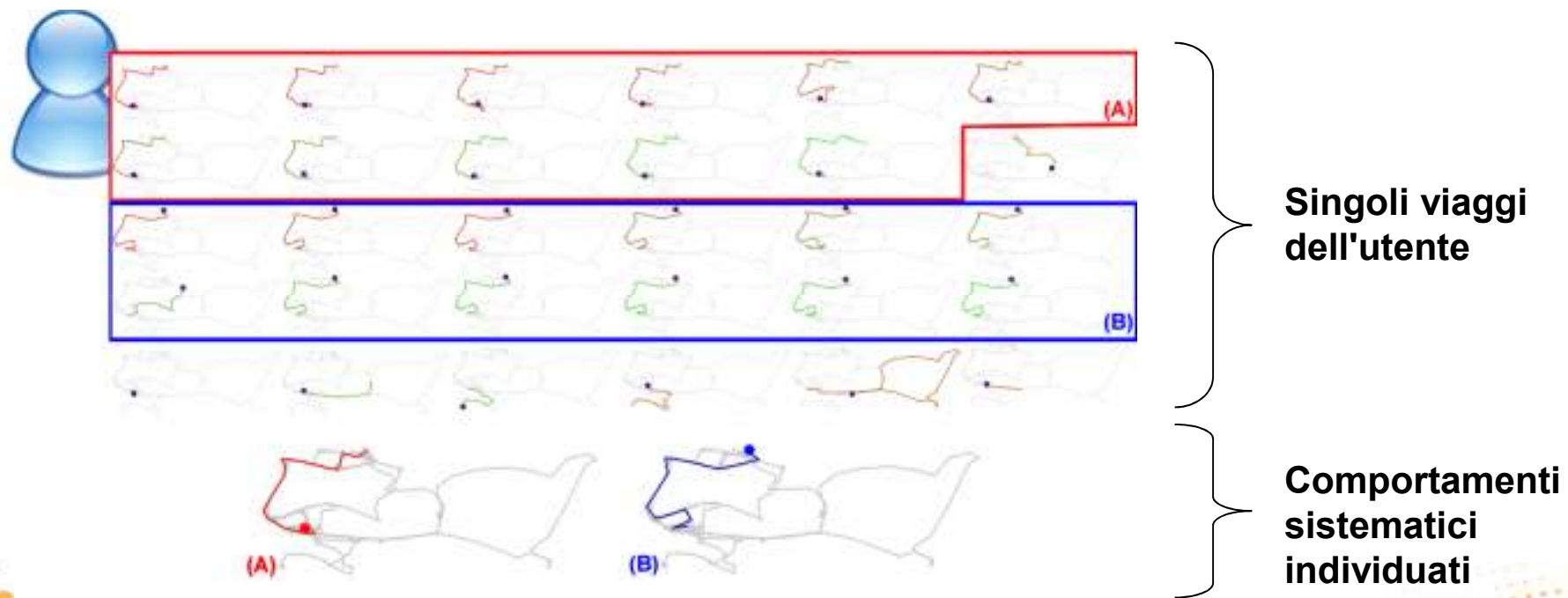
-
- How many commuters may use carpooling as an alternative to their own cars?



Individual travellers' profiles



Mobility routines



Car pooling potential

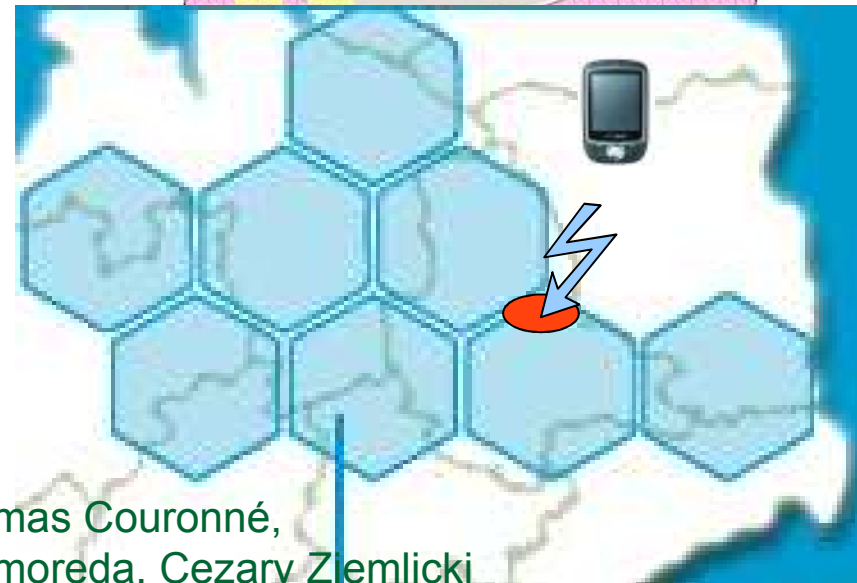
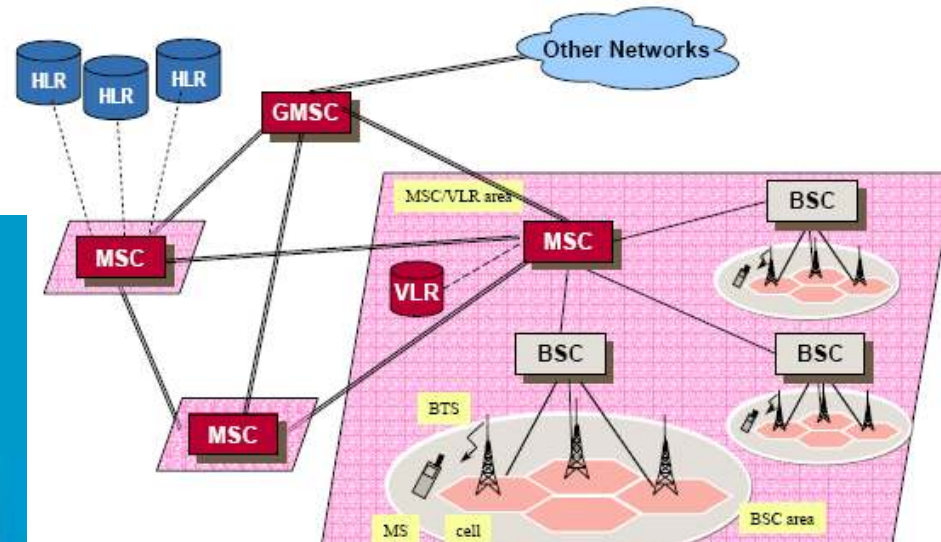
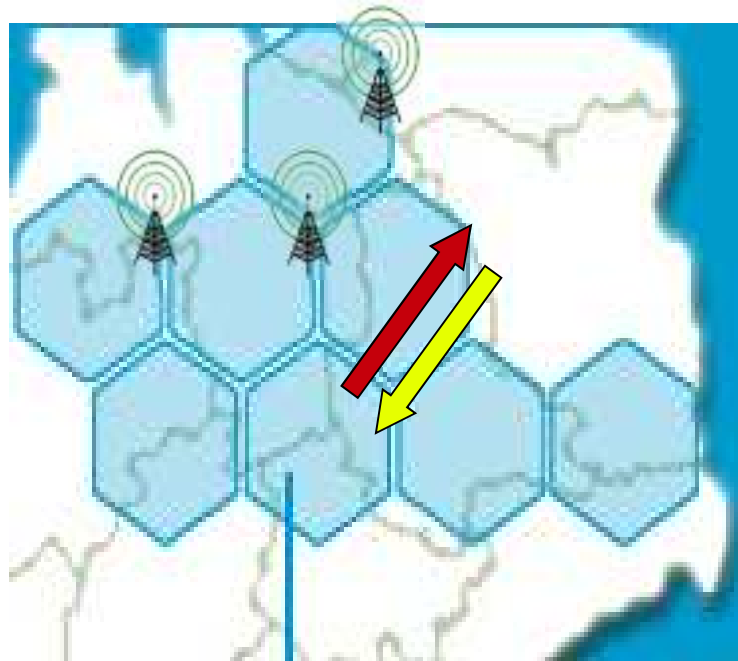
32.5% users share one or more routines with other users



■ How do tourists visit Paris?



GSM roaming CDR data – Orange Paris

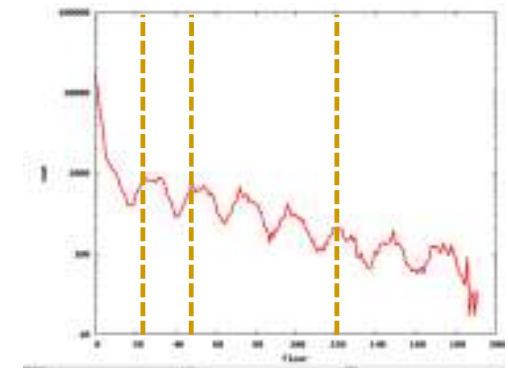
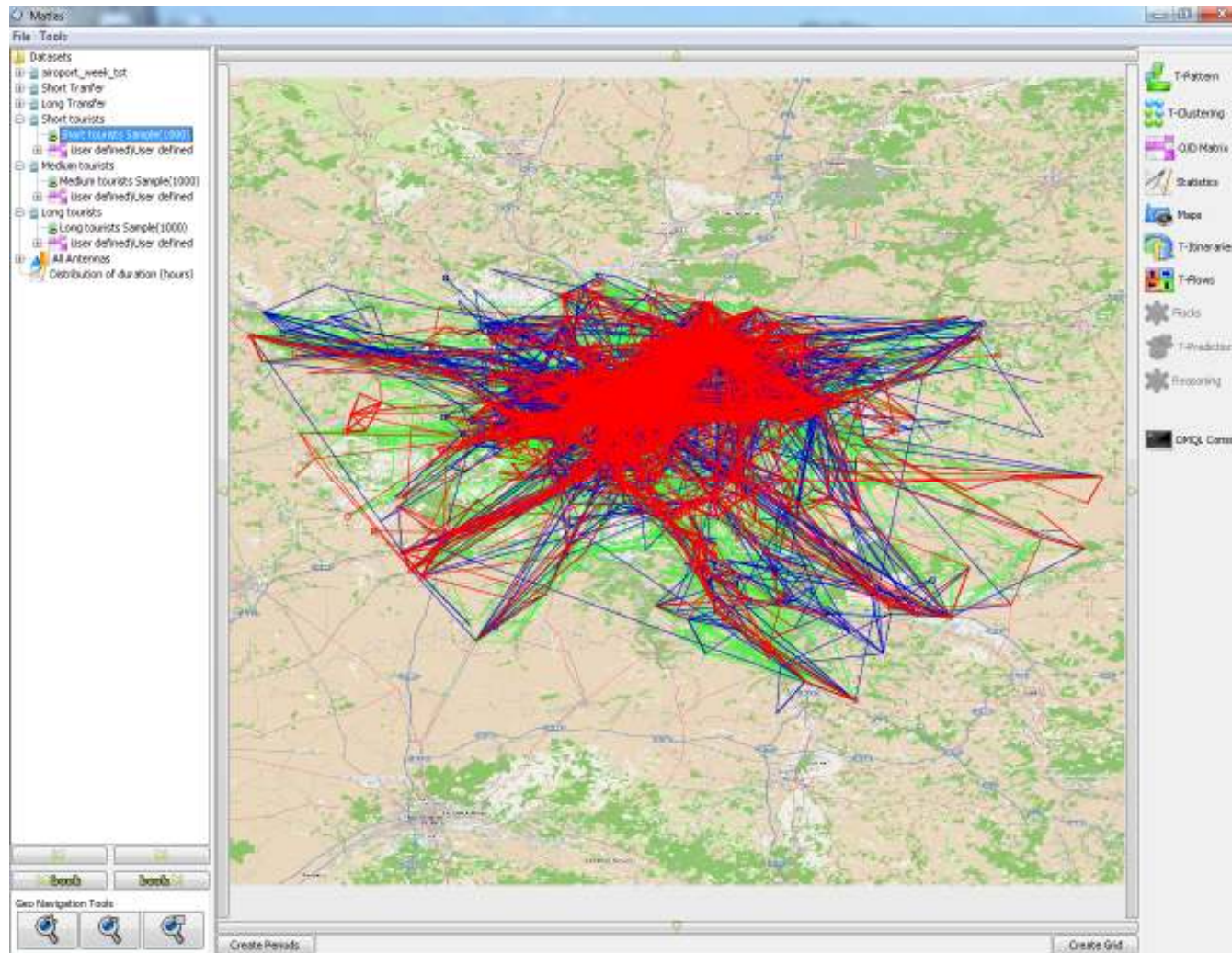


Ana-Maria Olteanu, Roberto Trasarti, Thomas Couronné,
Fosca Giannotti, Mirco Nanni, Zbigniew Smoreda, Cezary Ziemlicki

GSM data analysis for tourism

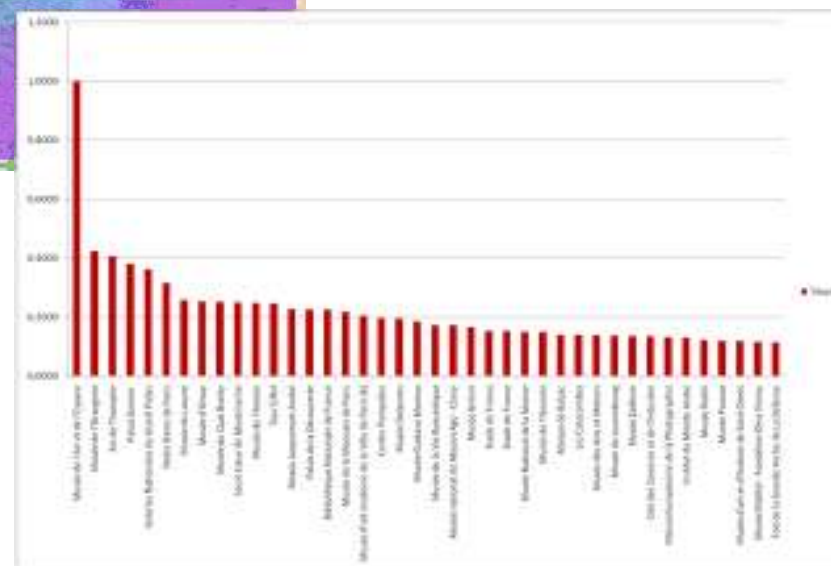
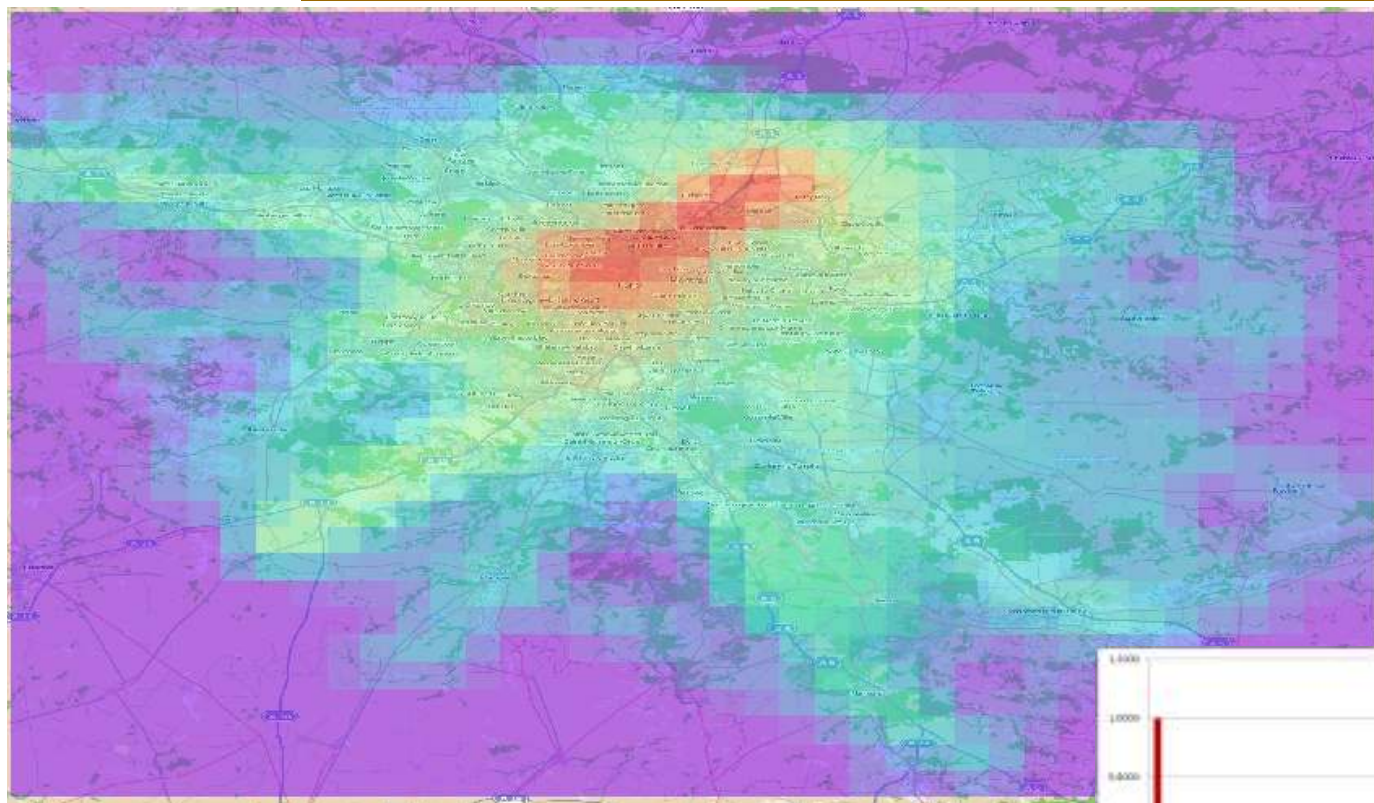
7th Int. Symp. on Spatial Data Quality (ISSDQ 2011)

Tourist visitation patterns

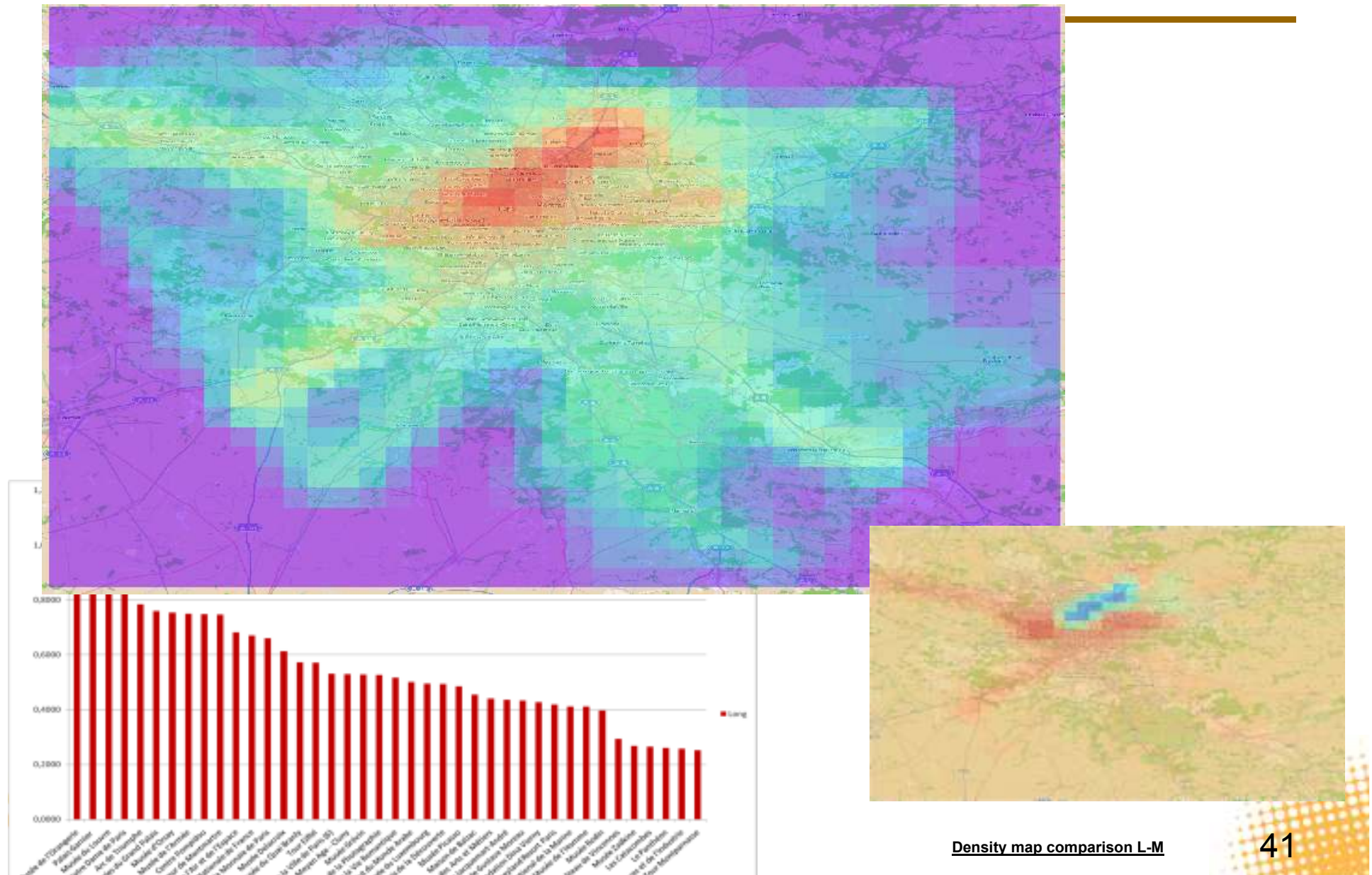


- Short visits 1 or 2 days
- Medium visits between 3 and 5 days
- Long visits more than 6 days

Short visits



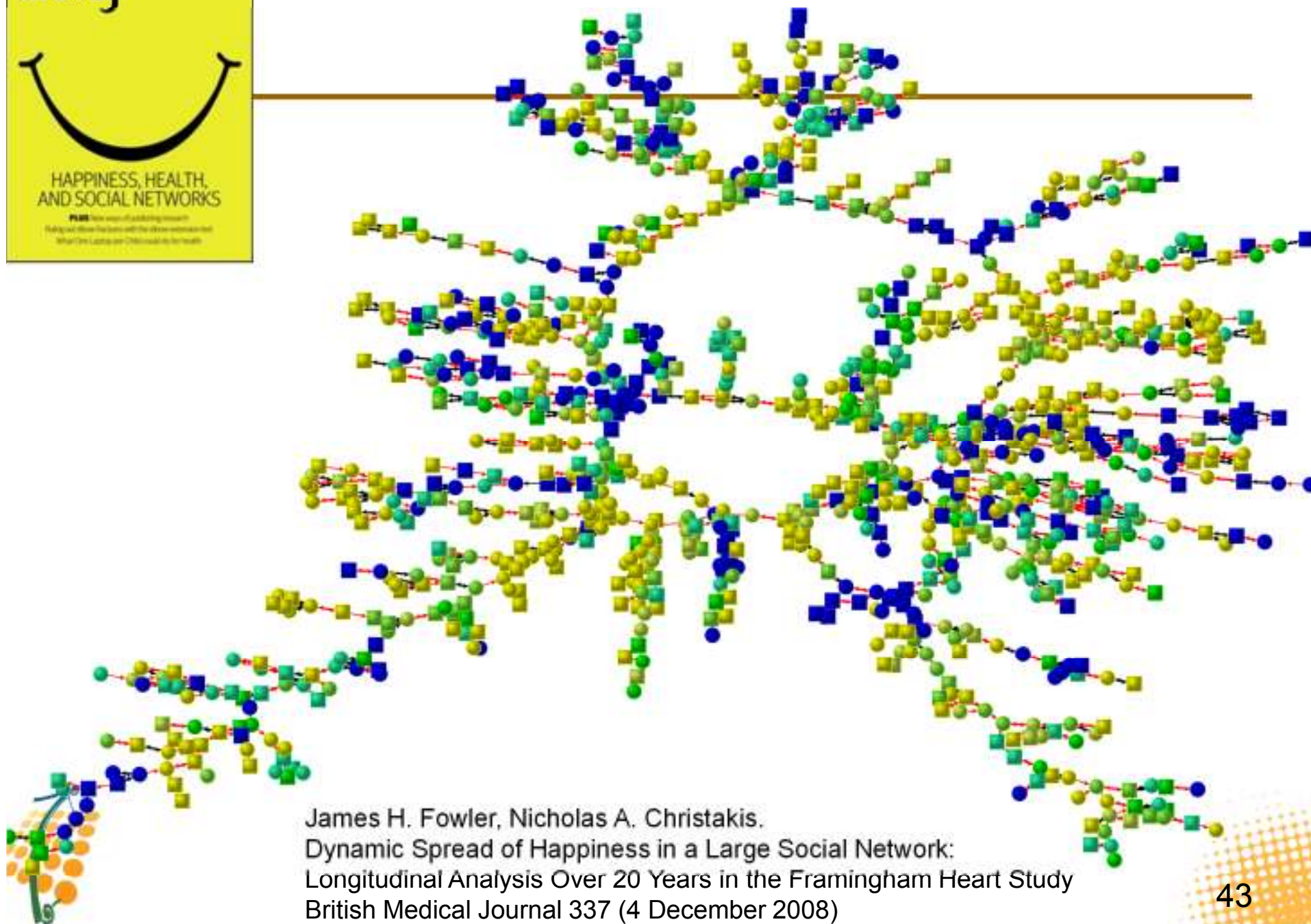
Long Visits



Density map comparison L-M

-
- What is your chance to become happy depending on the level of happiness in your social network?





James H. Fowler, Nicholas A. Christakis.
Dynamic Spread of Happiness in a Large Social Network:
Longitudinal Analysis Over 20 Years in the Framingham Heart Study
British Medical Journal 337 (4 December 2008)

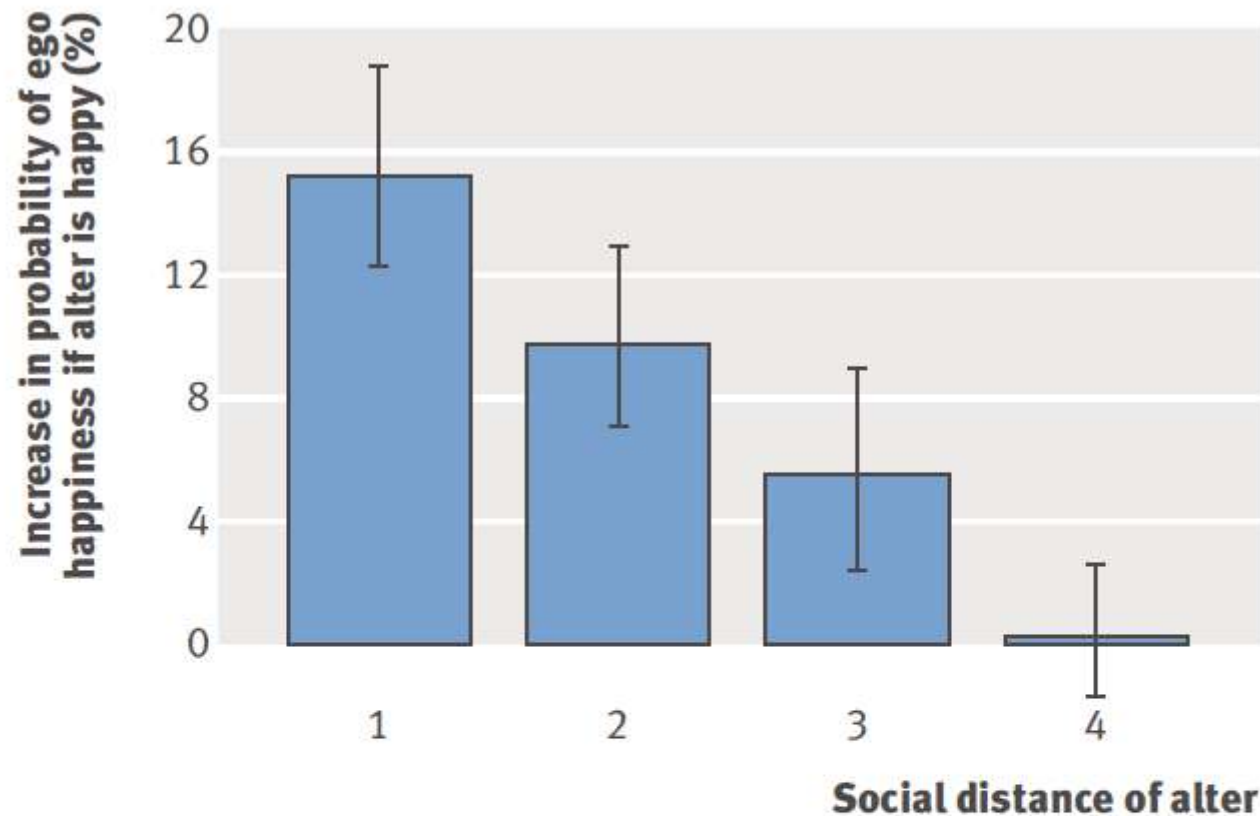
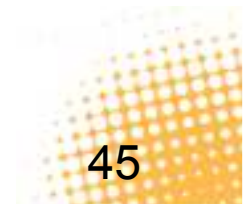


Fig 2 | Social distance and happiness in the Framingham social network. Percentage increase in likelihood an ego is happy if friend or family member at certain social distance is happy (instead of unhappy). The relationship is strongest between individuals who are directly connected but remains significantly >0 at social distances up to three degrees of separation, meaning that a person's happiness is associated with happiness of people up to three degrees removed from them in the network. Values derived by comparing conditional

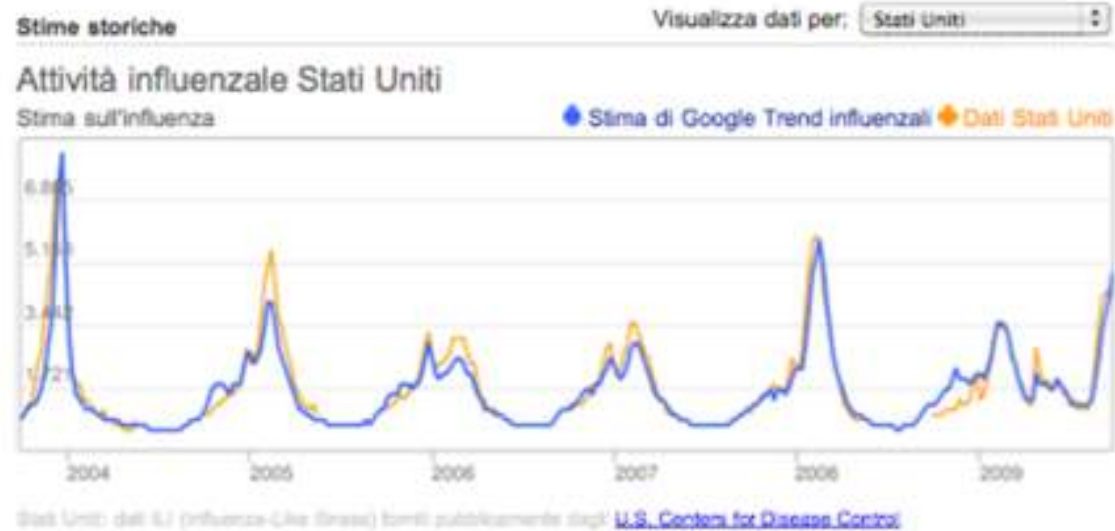


Comparison:
extra USD10,000
of annual income
has an average
effect of 2%

-
- How to predict the spreading patterns of the next flu?



Google Flu Trends



Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

¹Google Inc. ²Centers for Disease Control and Prevention



Nature 457, 1012-1014 (2009)

Simulation of H1N1 virus spreading

- <http://vimeo.com/17840291>
- Alex Vespignani and colleagues at Indiana University & ISI Foundation, Torino



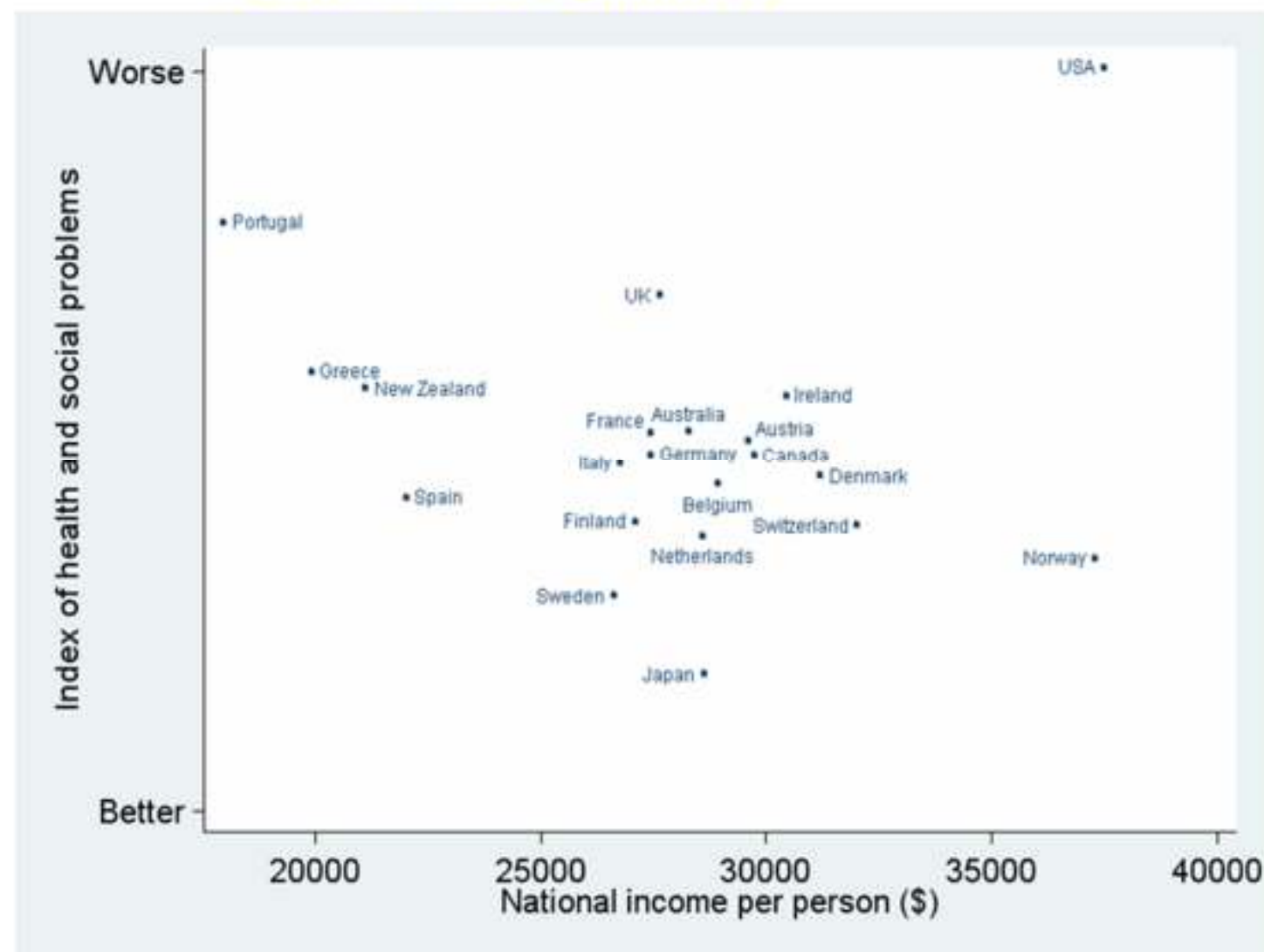
-
- What is an economical explanation of social well-being in the most developed nations?



Health and Social Problems are not Related to Average Income in Rich Countries

Index of:

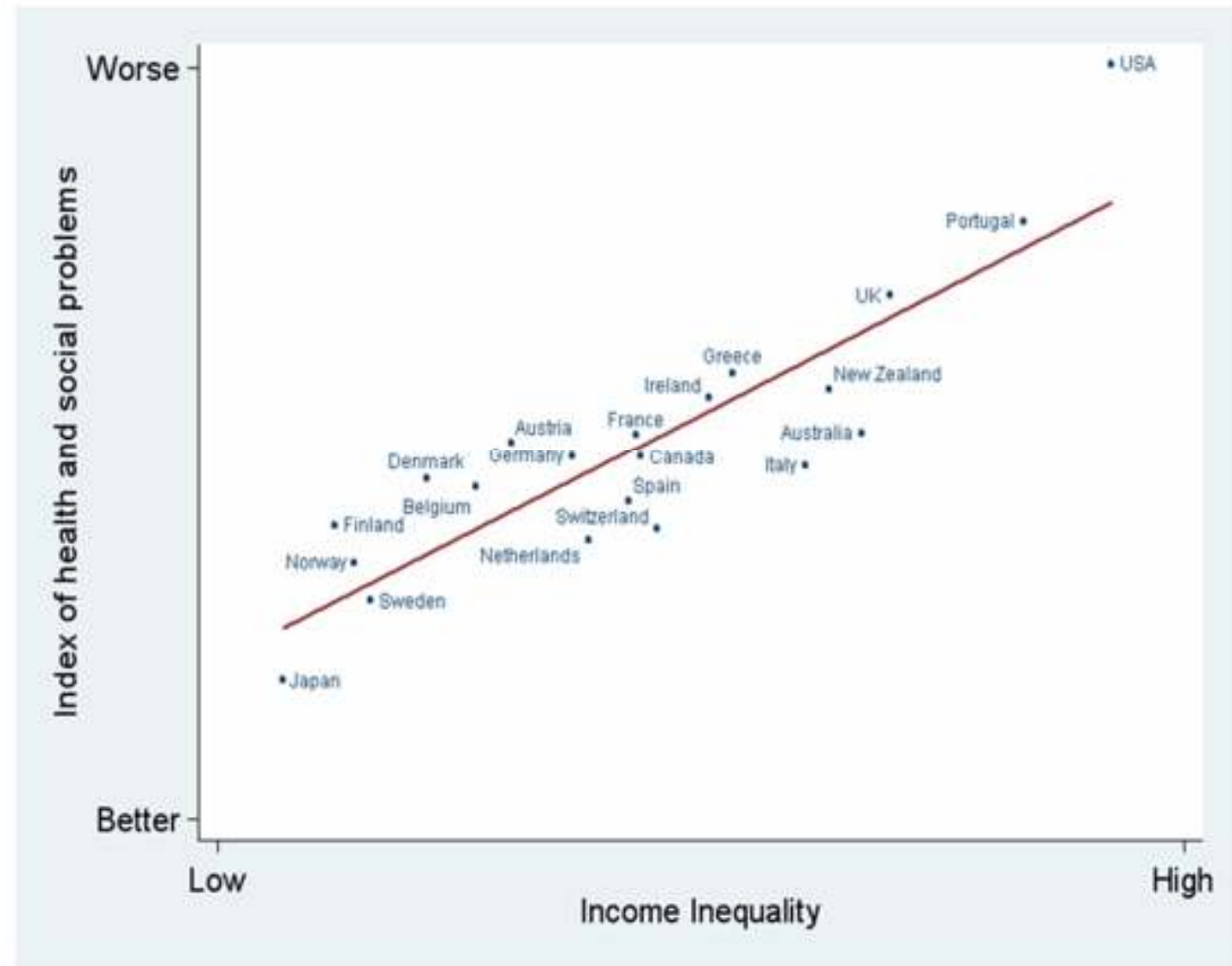
- Life expectancy
- Math & Literacy
- Infant mortality
- Homicides
- Imprisonment
- Teenage births
- Trust
- Obesity
- Mental illness – incl. drug & alcohol addiction
- Social mobility



Health and Social Problems are Worse in More Unequal Countries

Index of:

- Life expectancy
- Math & Literacy
- Infant mortality
- Homicides
- Imprisonment
- Teenage births
- Trust
- Obesity
- Mental illness – incl. drug & alcohol addiction
- Social mobility

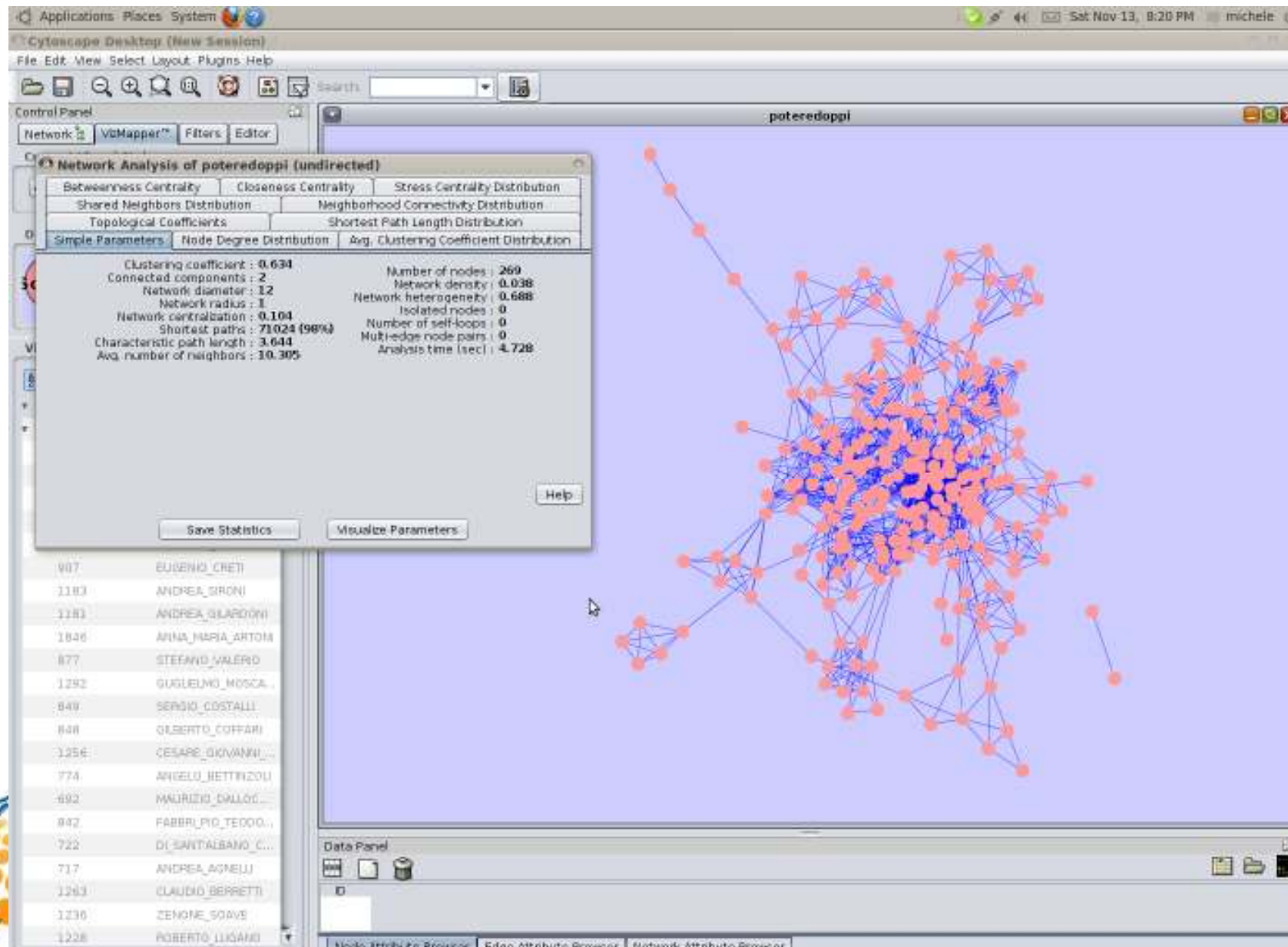


Source: Wilkinson & Pickett, *The Spirit Level* (2009)

-
- Which are the mains cliques or cartels in the network of managers sitting in multiple company boards?



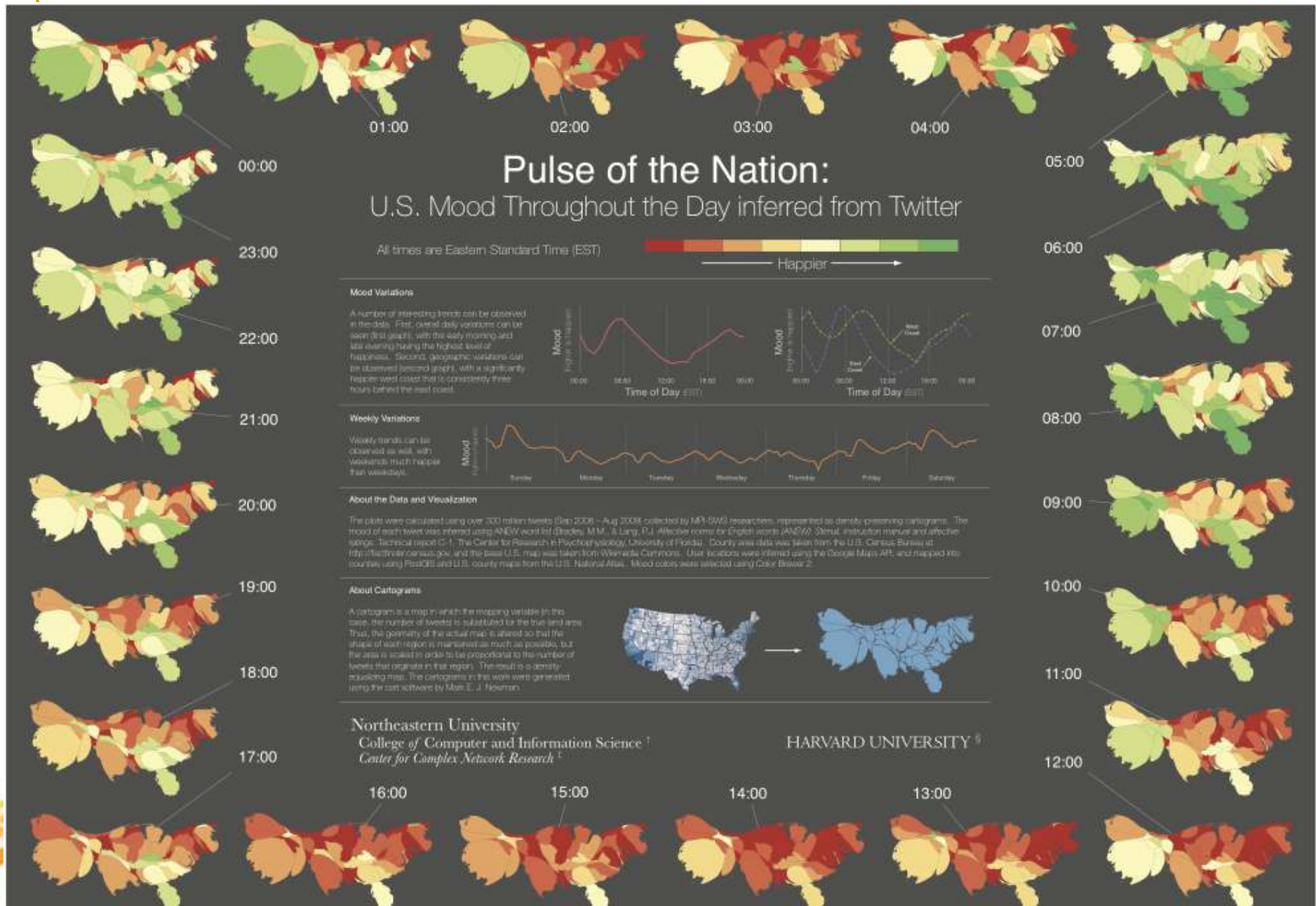
Map of Power



-
- How does an entire country's mood change around the clock?



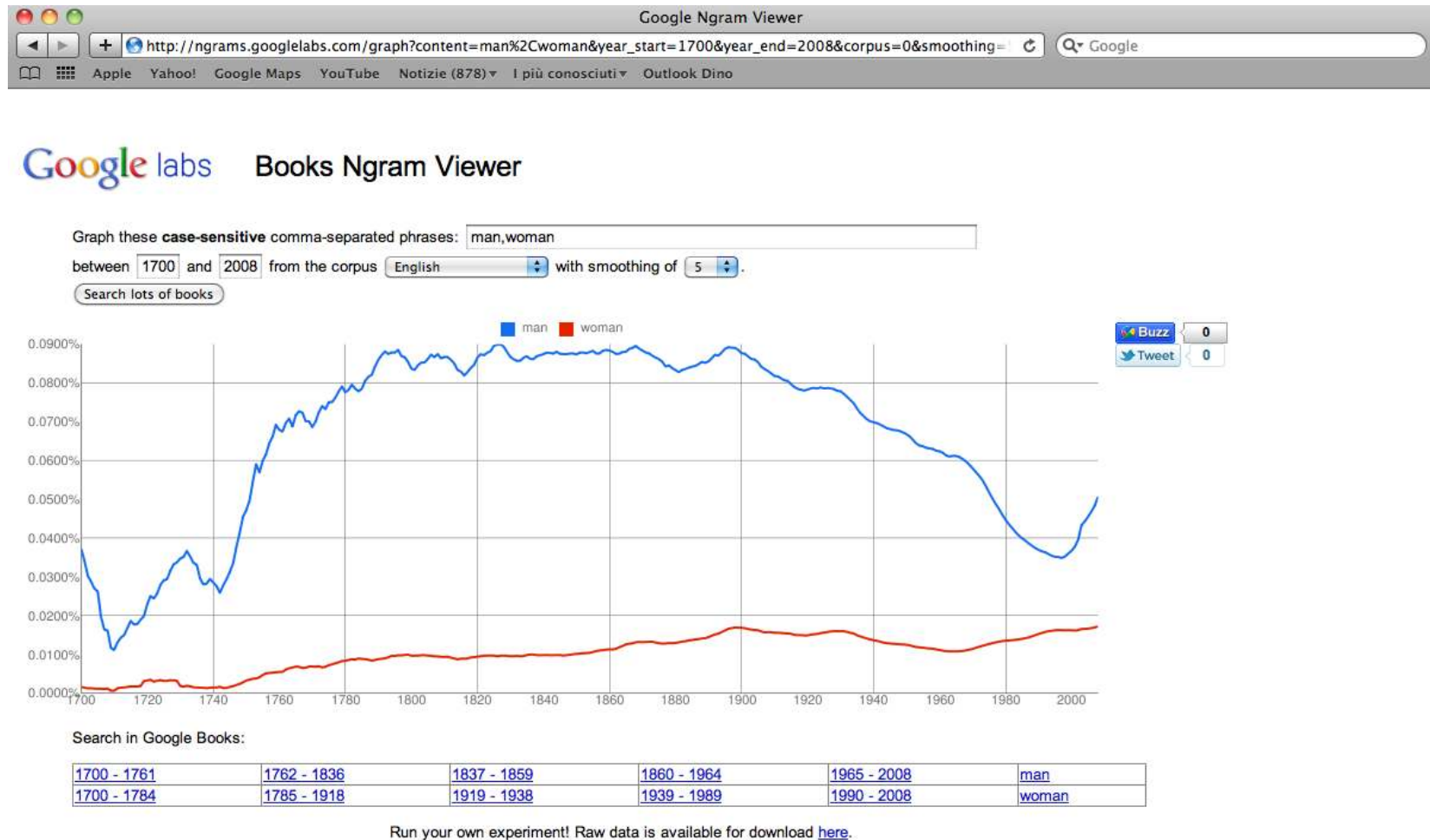
twittermood



-
- What is the impact of “man” and “woman” in the English literature (culture) from 1700 to nowadays?



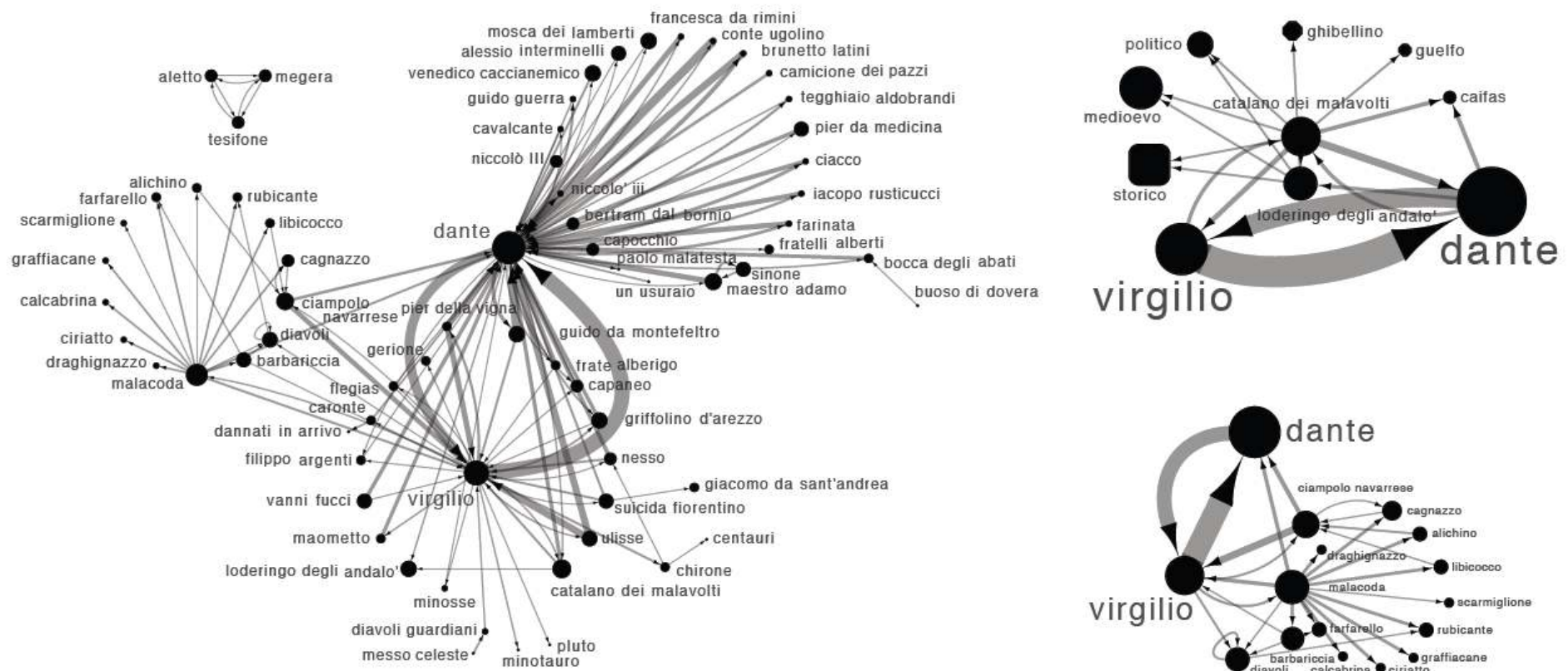
ngrams.googlelabs.com



*Quale è la rete sociale
dell'Inferno dantesco? Come
evolve nella narrazione?*



Fig. 1. Visualization of the *Inferno* social network, along with two conceptual ego networks, extracted from the main network and ontology data.



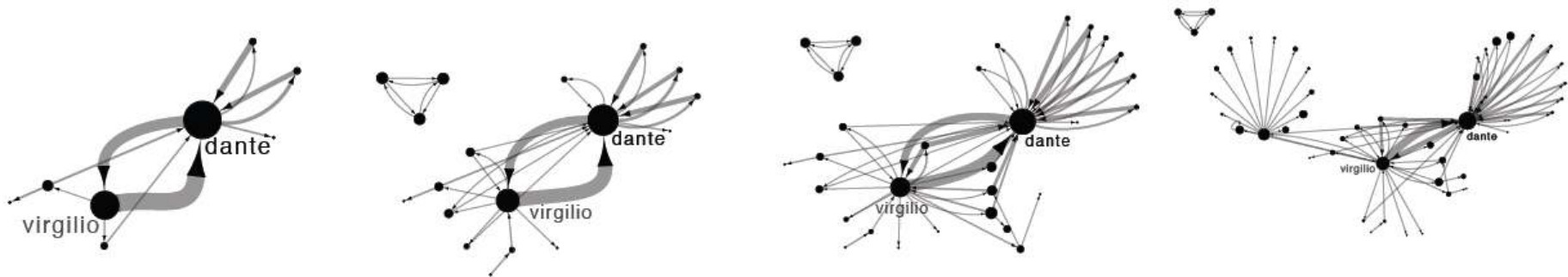


Fig. 2. Temporal evolution of the *Inferno* social network within the narrative timeline.

Cappelli, A., Coscia, M., Giannotti, F., Pedreschi, D. and Rinzivillo, S.
The Social Network of Dante's Inferno.

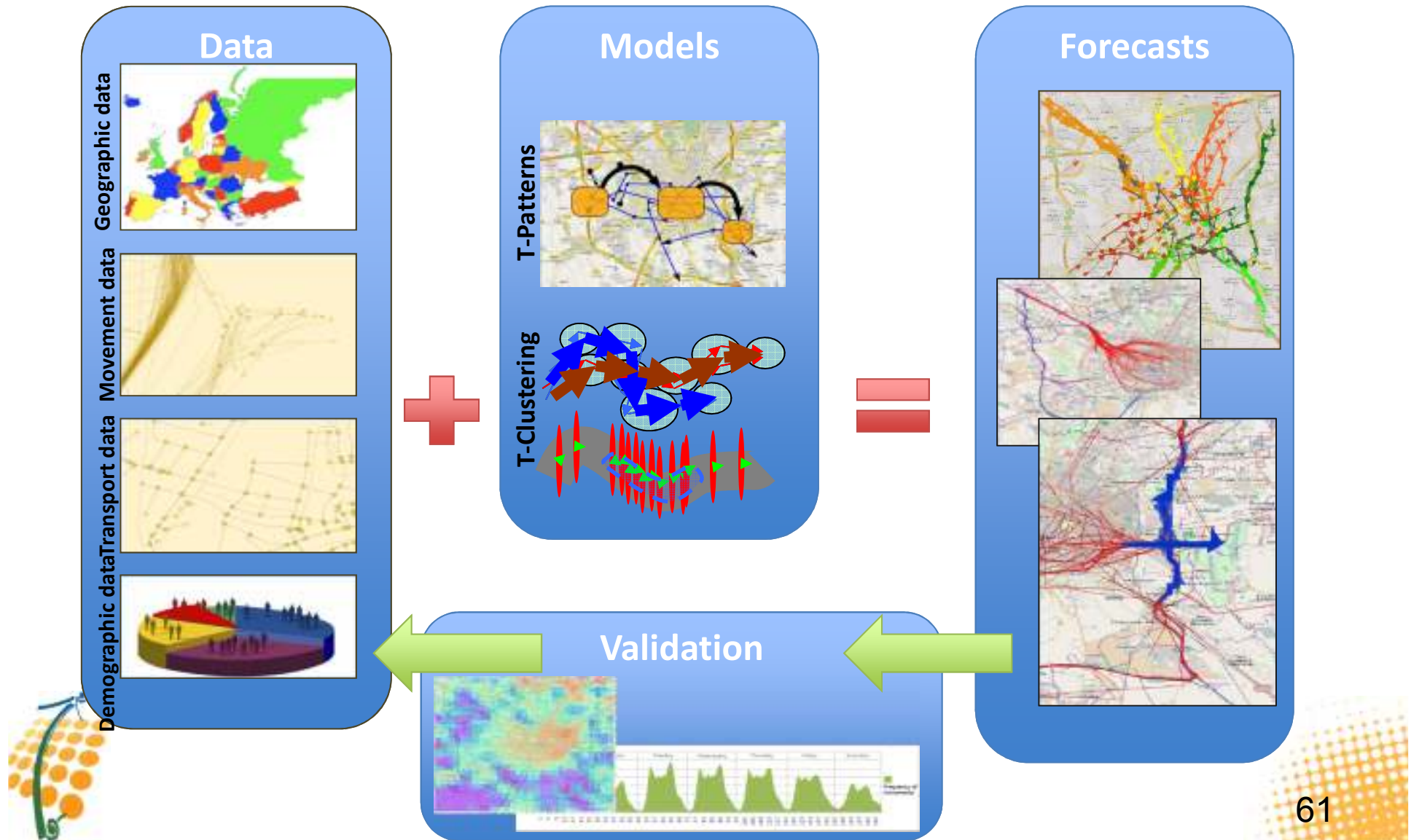
In Art, Humanities, Complex Networks – Leonardo Workshop at NetSci, 2010, Int. Conf. on Network Science.



What's behind the scene

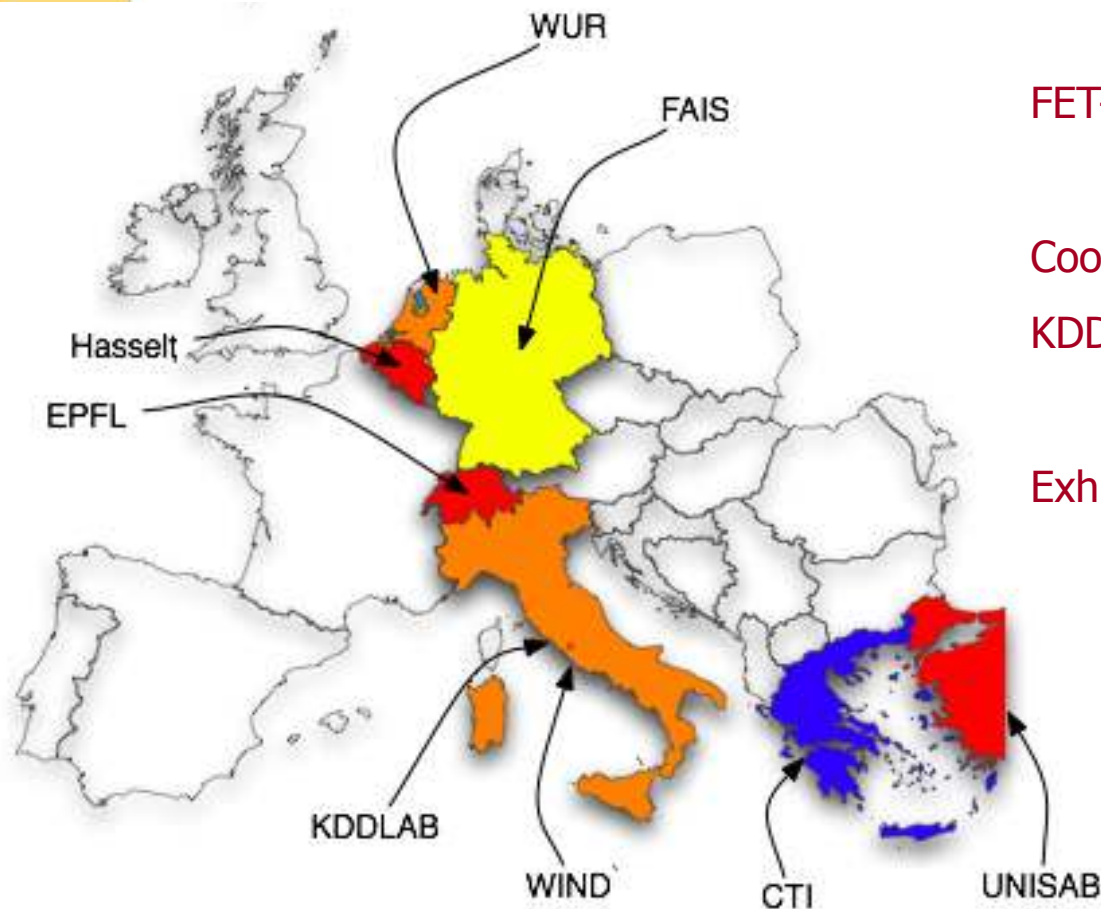


From DATA to KNOWLEDGE





www.geopkdd.eu

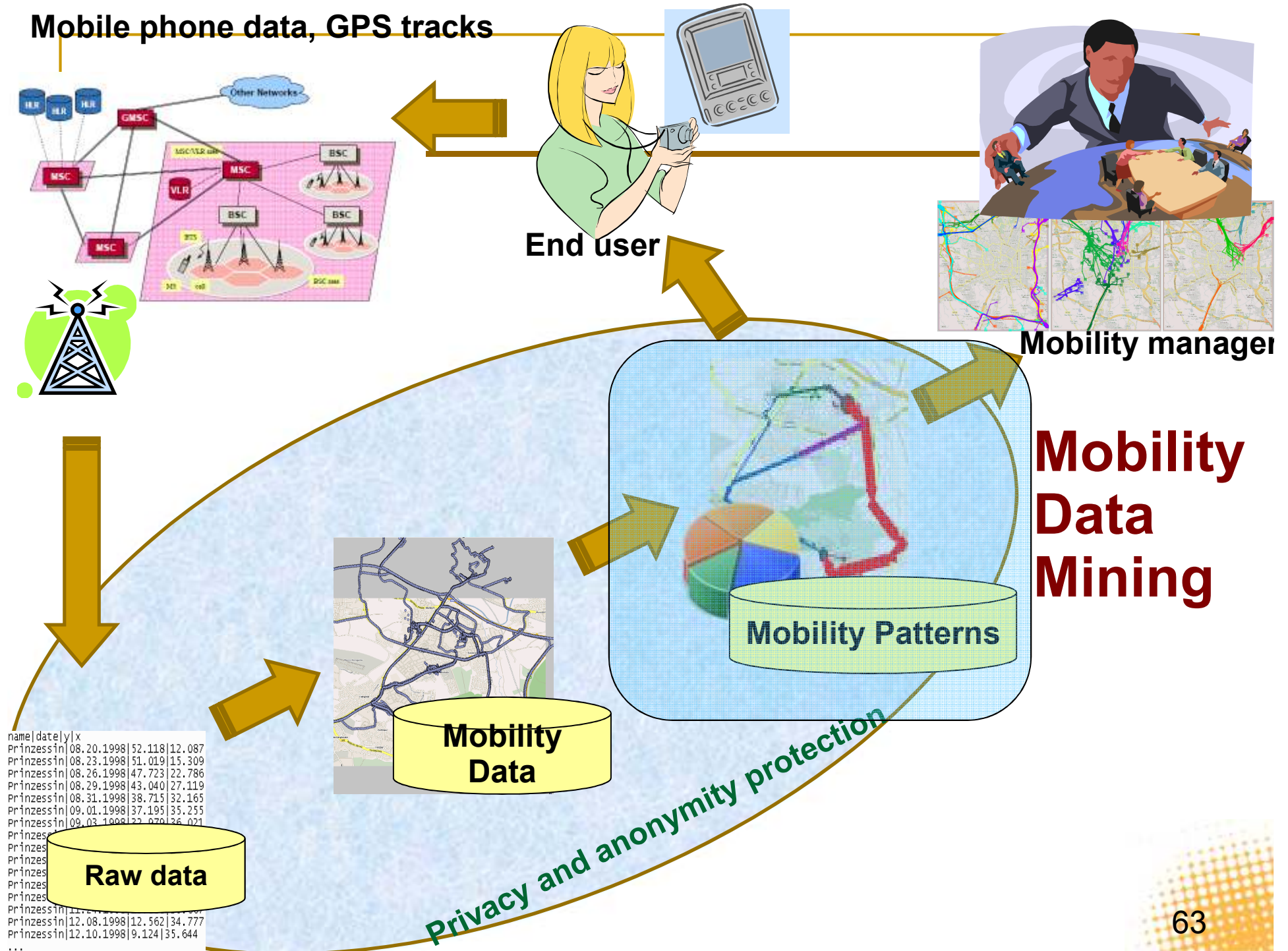


FET-Open STREP 2005-2009

Coordinator:
KDD-LAB Pisa, ISTI-CNR

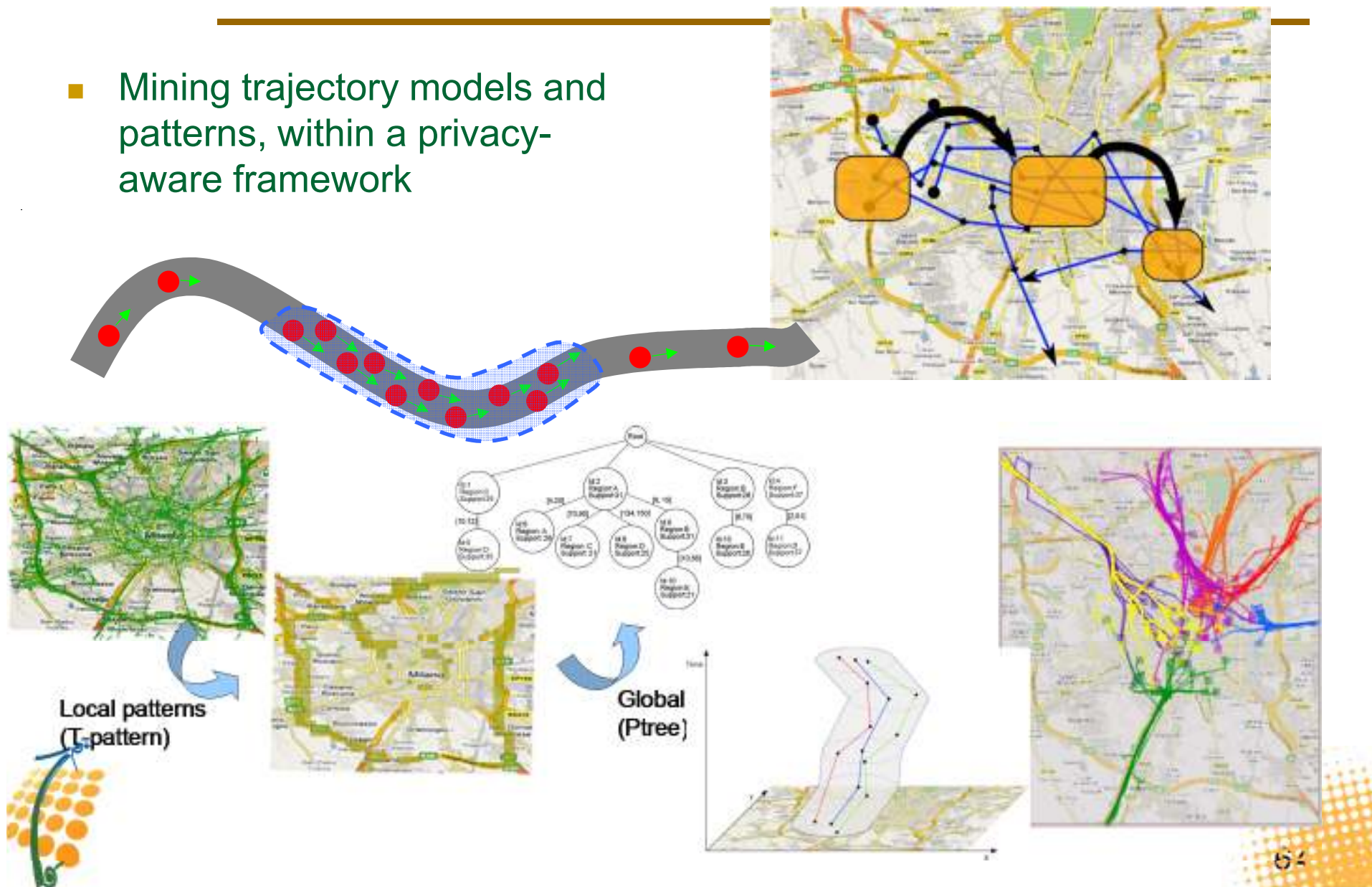
Exhibited at European Parliament
April 2010

Mobile phone data, GPS tracks



Mobility Data Mining

- Mining trajectory models and patterns, within a privacy-aware framework



Knowledge Discovery and Data Mining Laboratory

Web Site: <http://kdd.isti.cnr.it>

Personnel

Lab Head



Cappelli Amedeo



Giannotti Fosca



Pedreschi Dino



Turini Franco

Post Doc



Berlingerio Michele



Pinelli Fabio



Trasarti Roberto

Research Staff



Nanni Mirco



Renso Chiara



Rinzivillo Salvatore



Ruggieri Salvatore

PhD Student



Coscia Michele



Monreale Anna



Ong Rebecca



Pennacchioli Diego



Caterina D'angelo



Schifani Claudio



Falchi Chiara



Qu Zehui

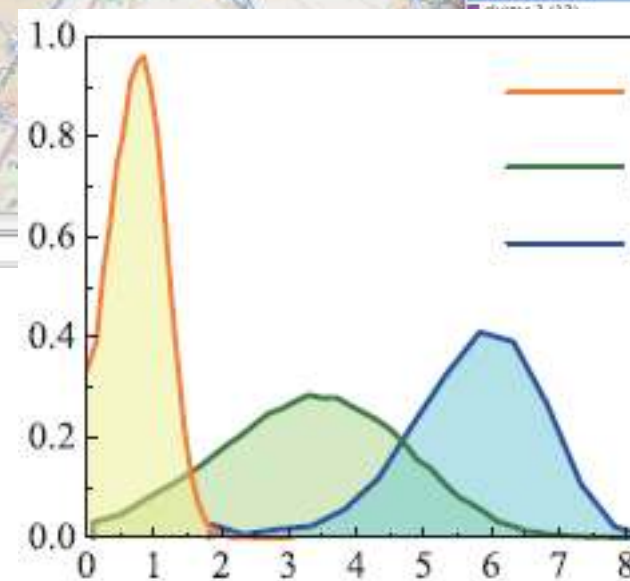
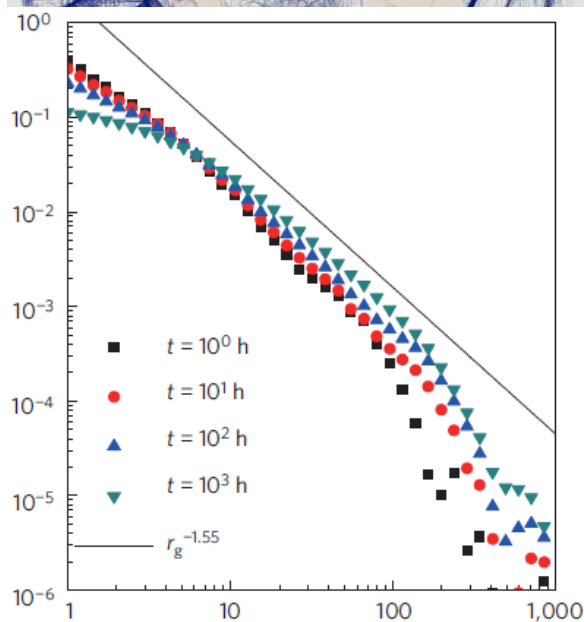
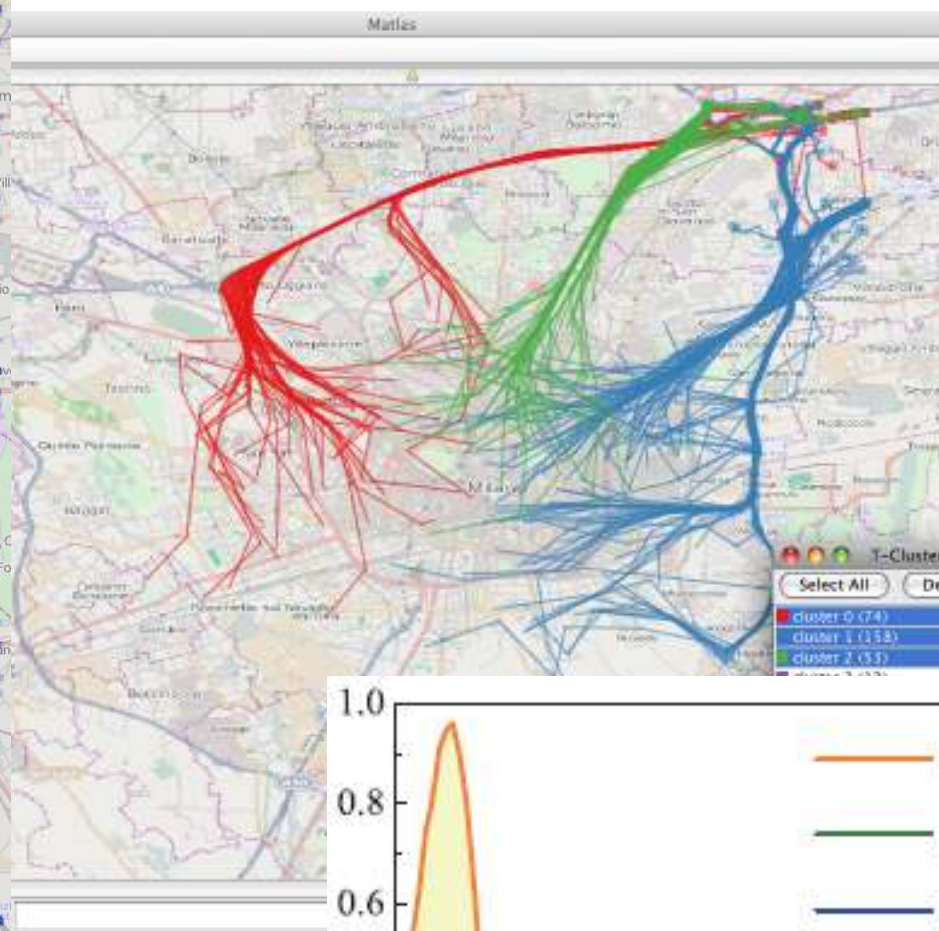
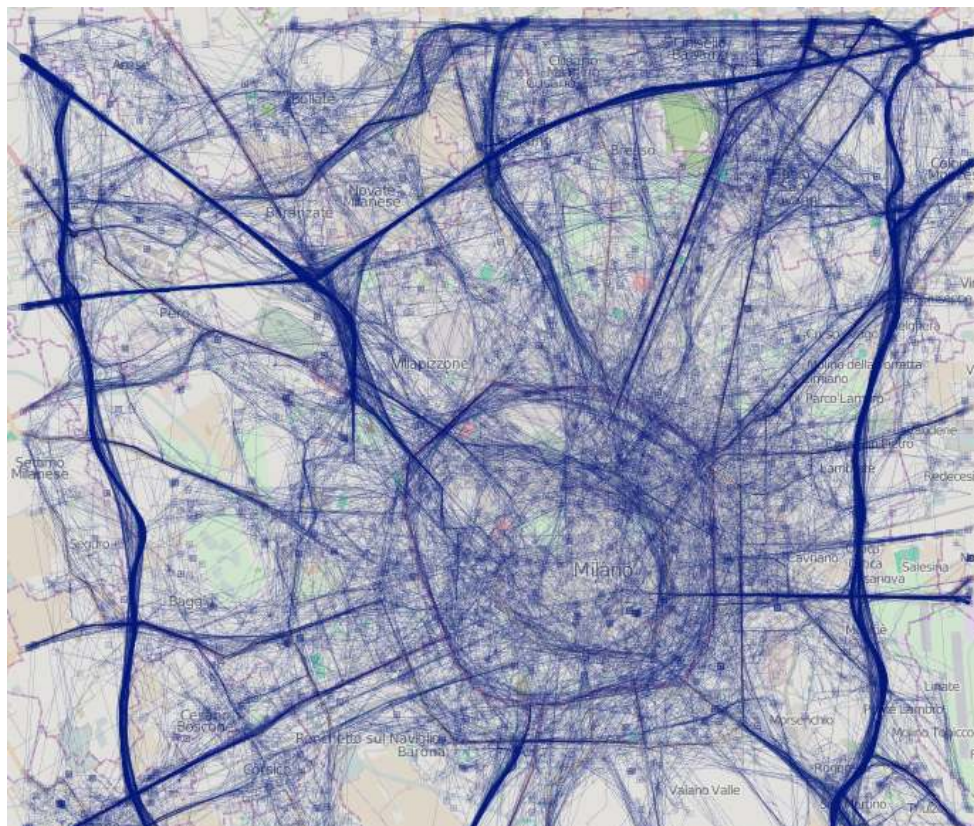


Furletti Barbara, Romei Andrea, Barsocchi Sergio

Towards a Social Data Science



social sciences + ICT + complexity science





FuturICT

www.futurict.eu

Aims of FuturICT

Bring together the fields of ICT, social science and complexity science

- to model our social system using
planetary scale simulations
- powered by a new
planetary scale data science
- to enable a paradigm shift to
data intensive governance

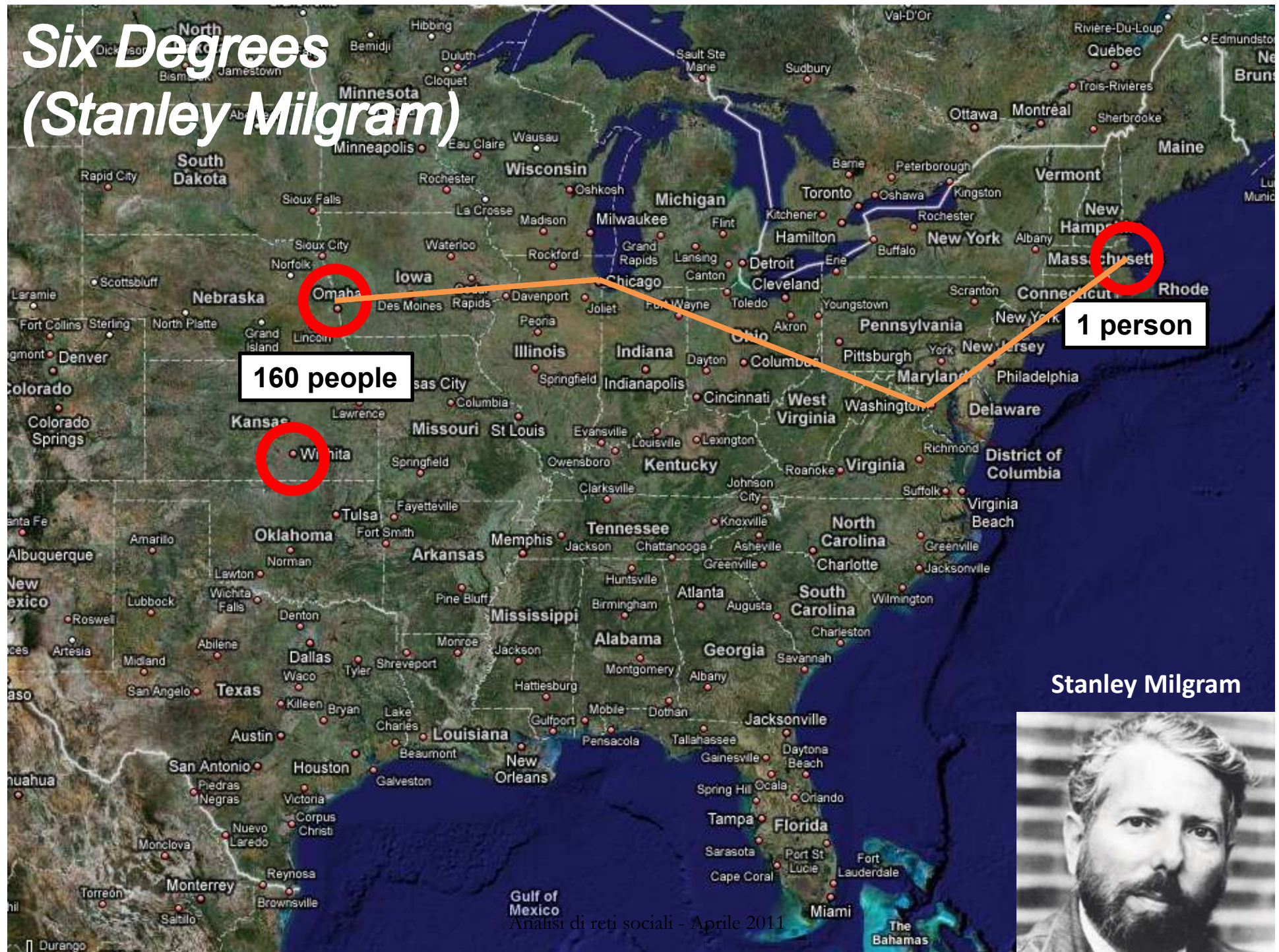
so we can manage our new global society in a sustainable manner.



*Early results: empirical
validation on big data of social
theories*



Six Degrees (Stanley Milgram)



Stanley Milgram



The Small-world experiment

- 64 chains completed:
 - 6.2 on the average, thus “6 degrees of separation”
- Further observations:
 - People who owned stock had shortest paths to the stockbroker than random people: 5.4 vs. 5.7
 - People from the Boston area have even closer paths: 4.4



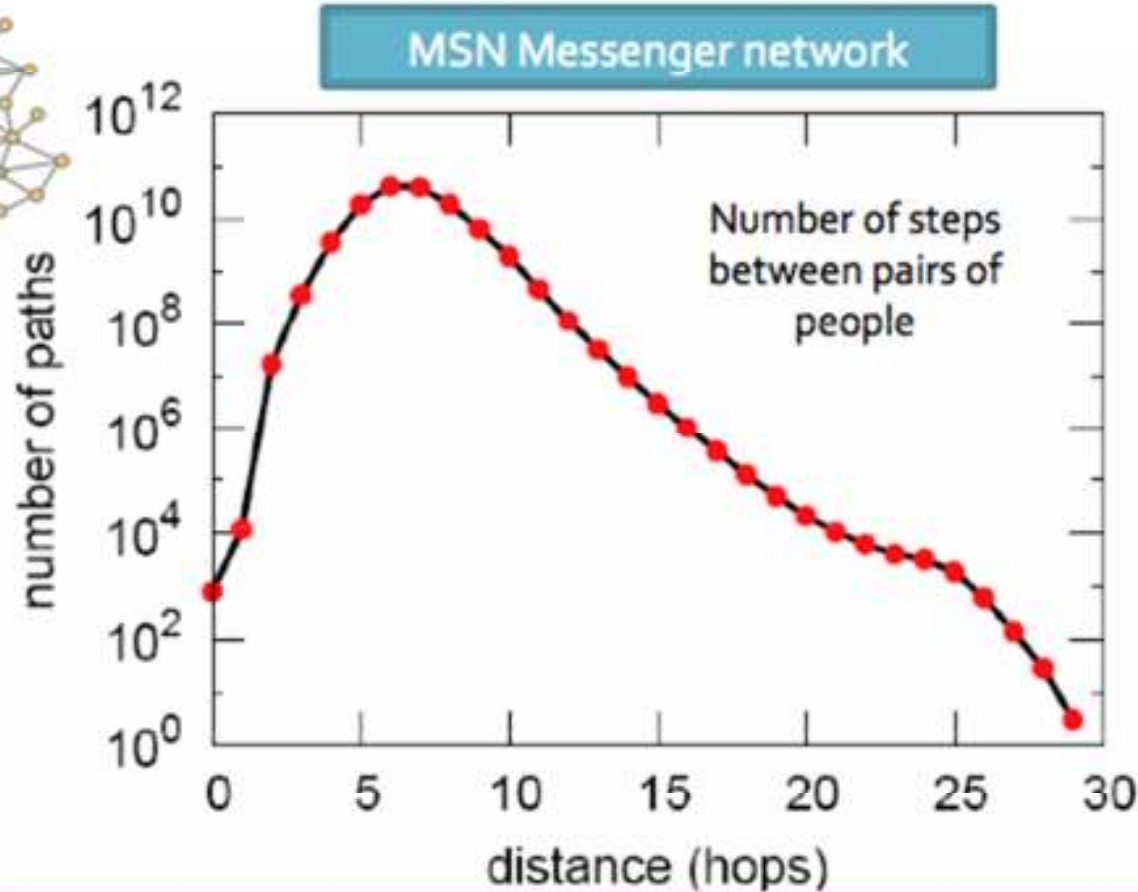
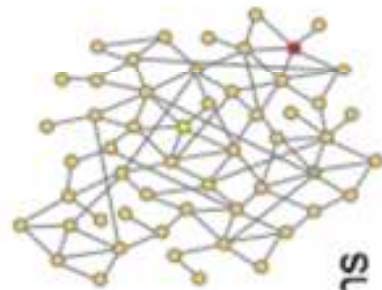
Conferma empirica dello small world

- Duncan Watts e colleghi, 2003, ripetono l'esperimento di Milgram usando l'email
 - 18 utenti target, 24.000 utenti di partenza
 - **Distanza media: 7**

- Jure Leskovec e colleghi, 2007, misurano la distanza media sul grafo di Messenger
 - 200 milioni di utenti
 - 30 miliardi di conversazioni
 - **Distanza media: 6.6**



MSN Network: Small world



Avg. path length 6.6
90% of the people can be reached in < 8 hops

Hops	Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

La forza dei legami deboli

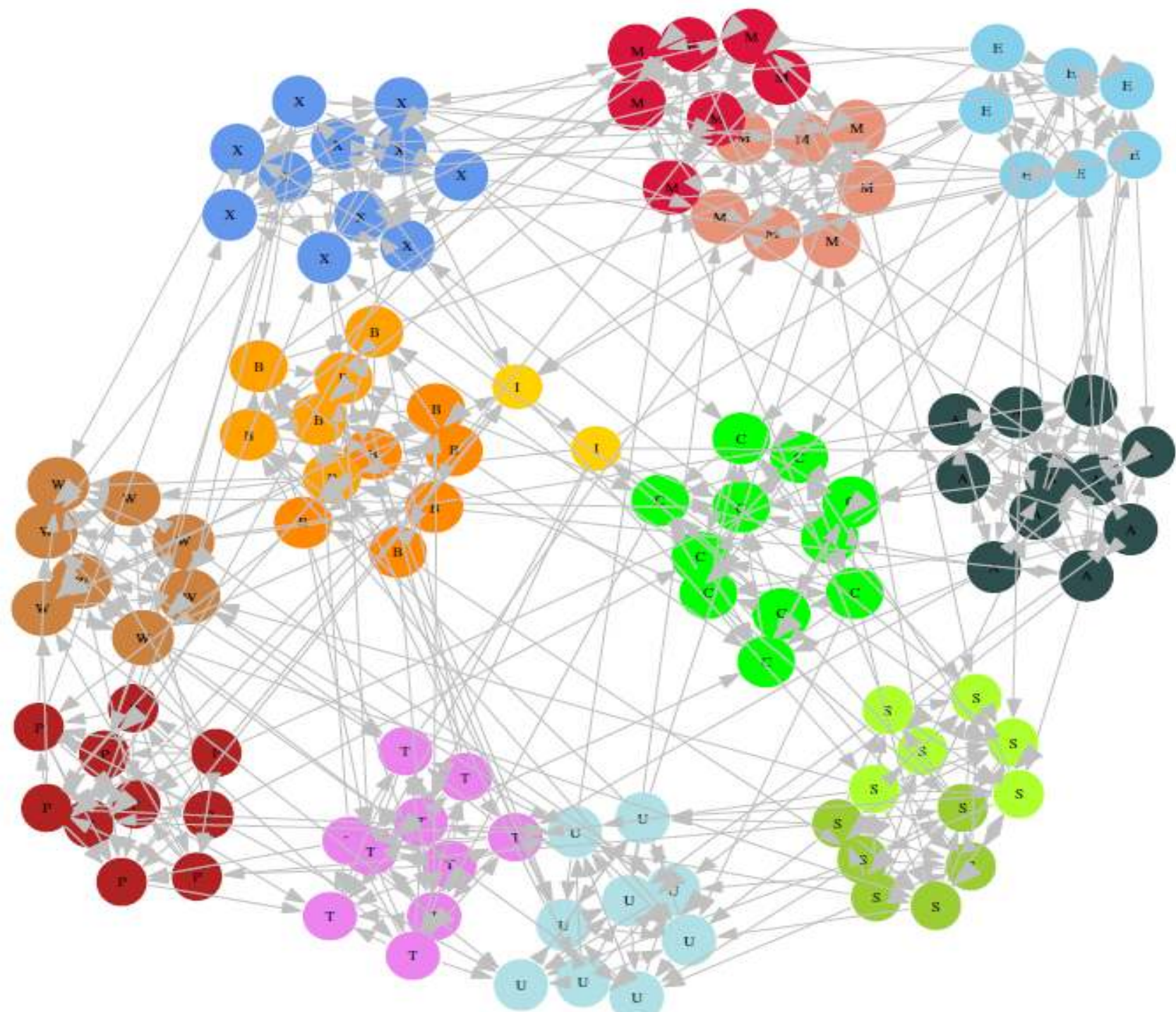
- Mark S. **Granovetter**, 1973
- Indagine: in che modo la gente viene a conoscenza di nuove opportunità di lavoro?
- Attraverso contatti personali
- Sorpresa: molto spesso **non** amici stretti, ma conoscenze “alla lontana”
- Perché?

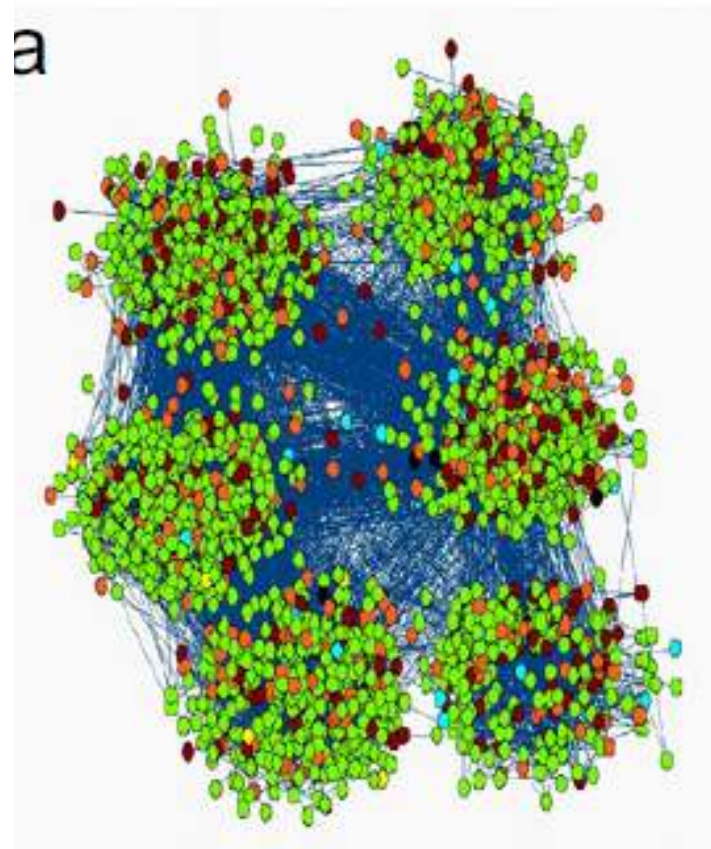
The Strength of Weak Ties

Mark S. Granovetter

American Journal of Sociology, Volume 78, Issue 6 (May, 1973), 1360-1380.







Node color

- Unknown
- Black
- Mixed
- Hispanic
- Asian
- White

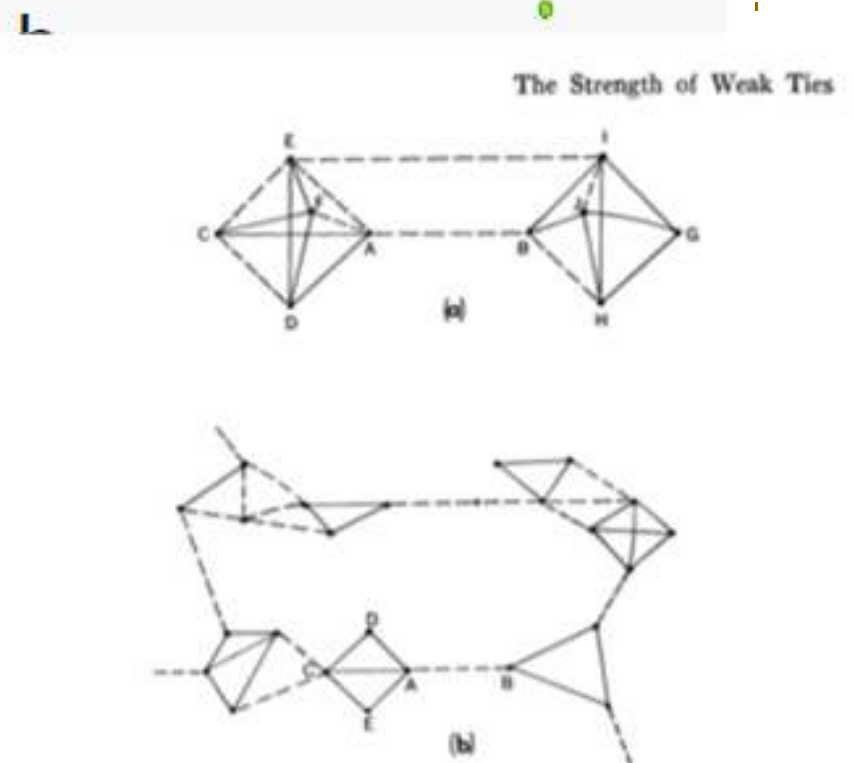
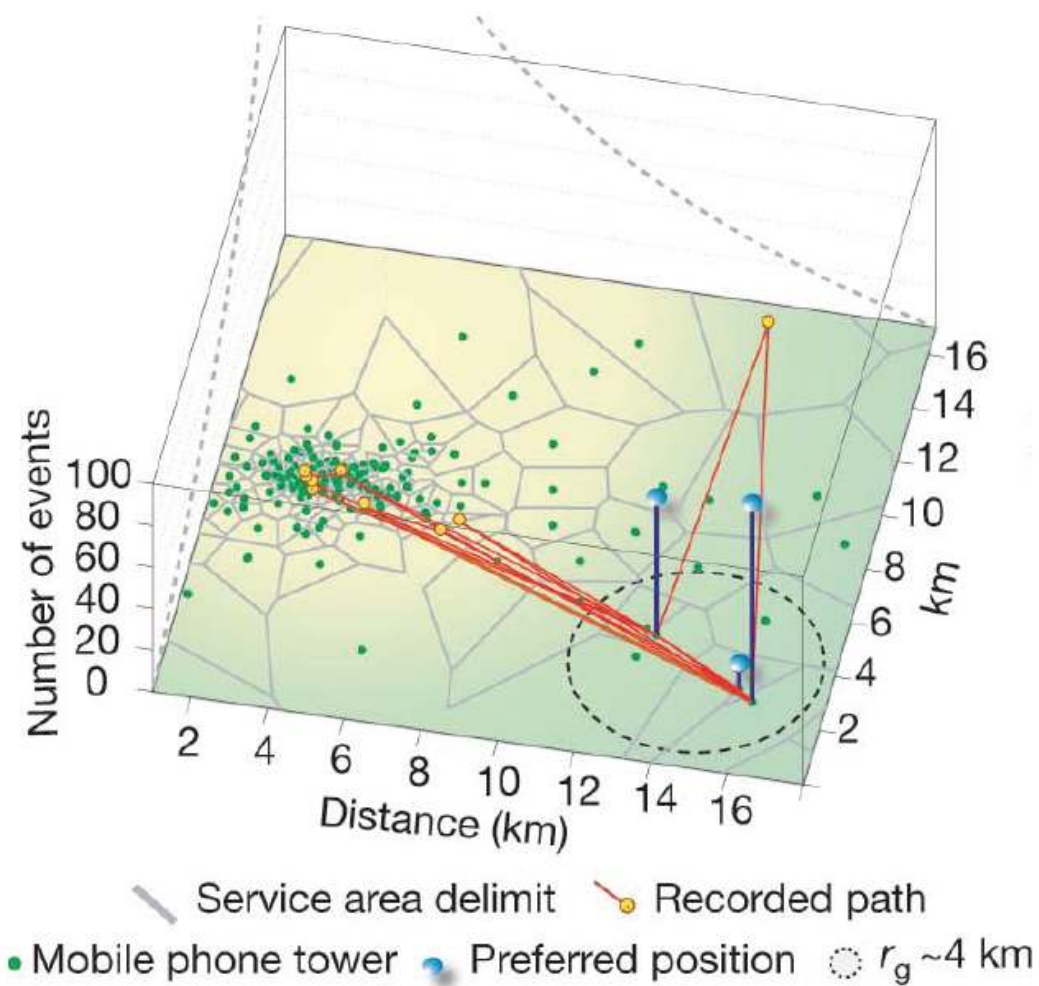


FIG. 2.—Local bridges. a, Degree 3; b, Degree 13. — = strong tie; --- = weak tie.



Country-wide tele-communication data



when
you
call



where
you
call



who
you
call

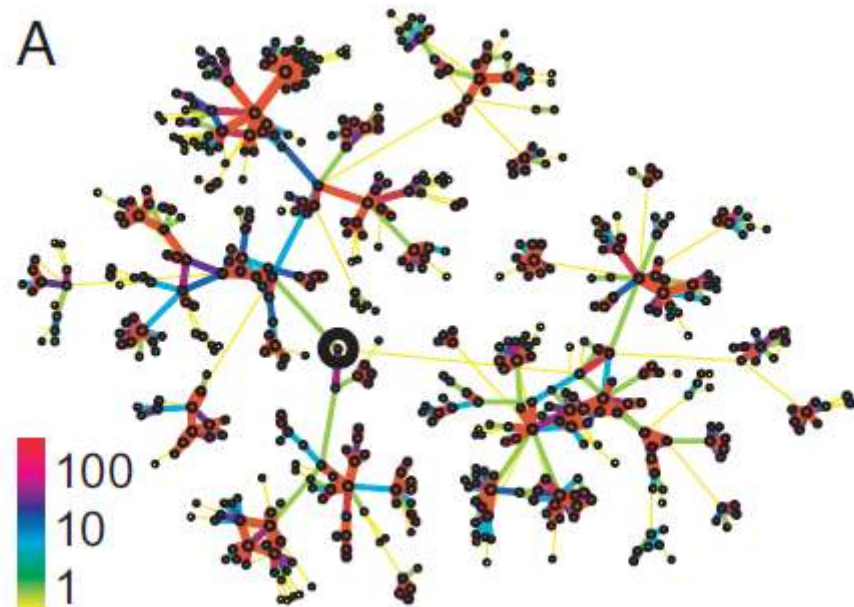
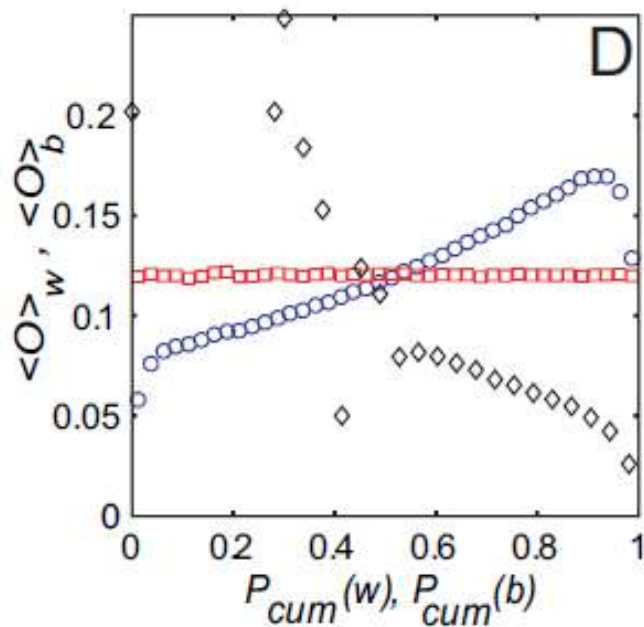
The social network

- I record della telefonia mobile sono un “proxy” delle interazioni sociali:
 - La rete “chi-chiama-chi”
- Diventa possibile la validazione su larga scala delle teorie delle scienze sociali
 - Vicinanza sociale di due utenti = numero degli amici comuni
 - Forza del legame sociale di due utenti = numero di chiamate



Validazione dell'ipotesi di Granovetter

- Onnela et al. PNAS 2007



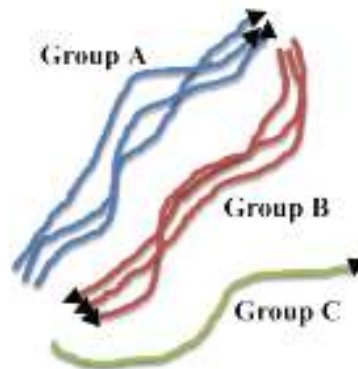
Analisi delle reti sociali

- Come scoprire le **comunità**?
- Quali sono le persone (nodi) più influenti (**centrali**)?
- Come si **diffonde** l'informazione (o una epidemia)?
- Come **evolve** nel tempo le rete? Come simularne la crescita?
- Quali **nuovi legami** si formano? E' possibile prevederli?
- Come interagiscono i **legami sociali** e la **mobilità**?

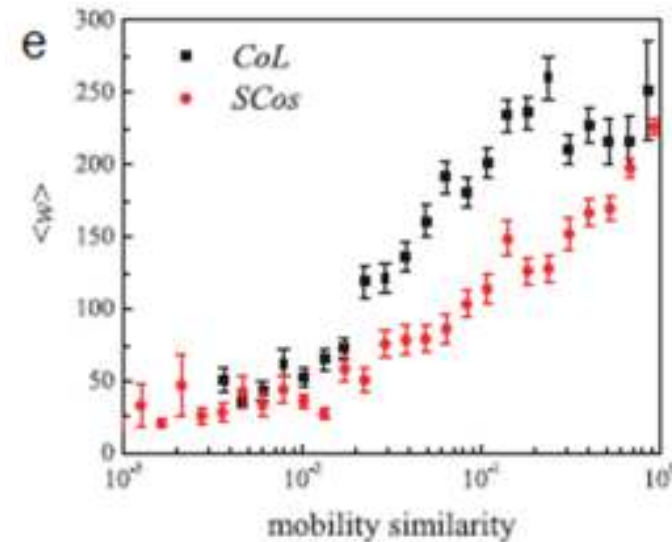
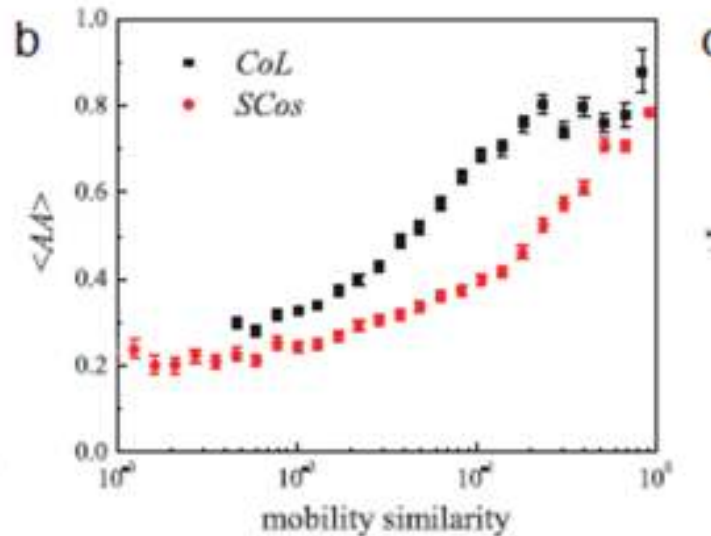


Quanto è simile la mobilità di due persone?

- E' possibile misurare quanto differiscono le traiettorie di ogni coppia di persone:
 - probabilità di incontrarsi
 - ovvero, quanto spesso si trovano contemporaneamente nello stesso posto



Colocation and social proximity are correlated

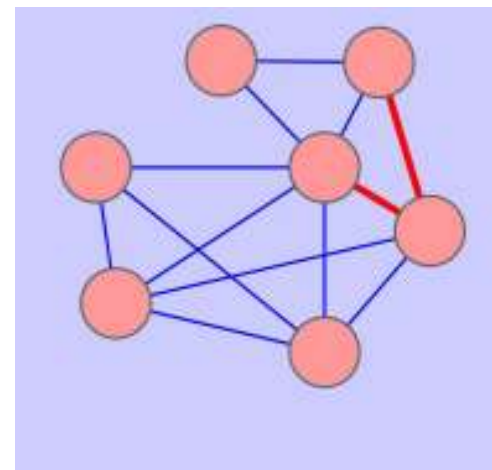
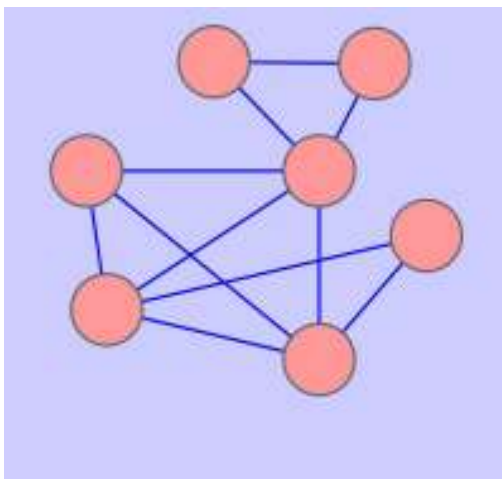


- co-location, network proximity and tie strength strongly correlate with each other
- measured on 3 months of calls, 6 Million users
- Wang, Pedreschi, Song, Giannotti, Barabási: Human mobility, social ties and link prediction. KDD 2011



Link prediction

- Come prevedere i nuovi link che si formeranno?



Link prediction, mobilità e legami sociali

- Probabilità di prevedere correttamente un nuovo link scegliendo una coppia di utenti a caso:
 - 10^{-6}
- Probabilità usando solo informazioni sociali (amici a comune):
 - 10%
- Probabilità usando sia informazioni sociali che mobilità:
 - **75%**





Quanto è prevedibile il movimento umano?



Prevedibilità

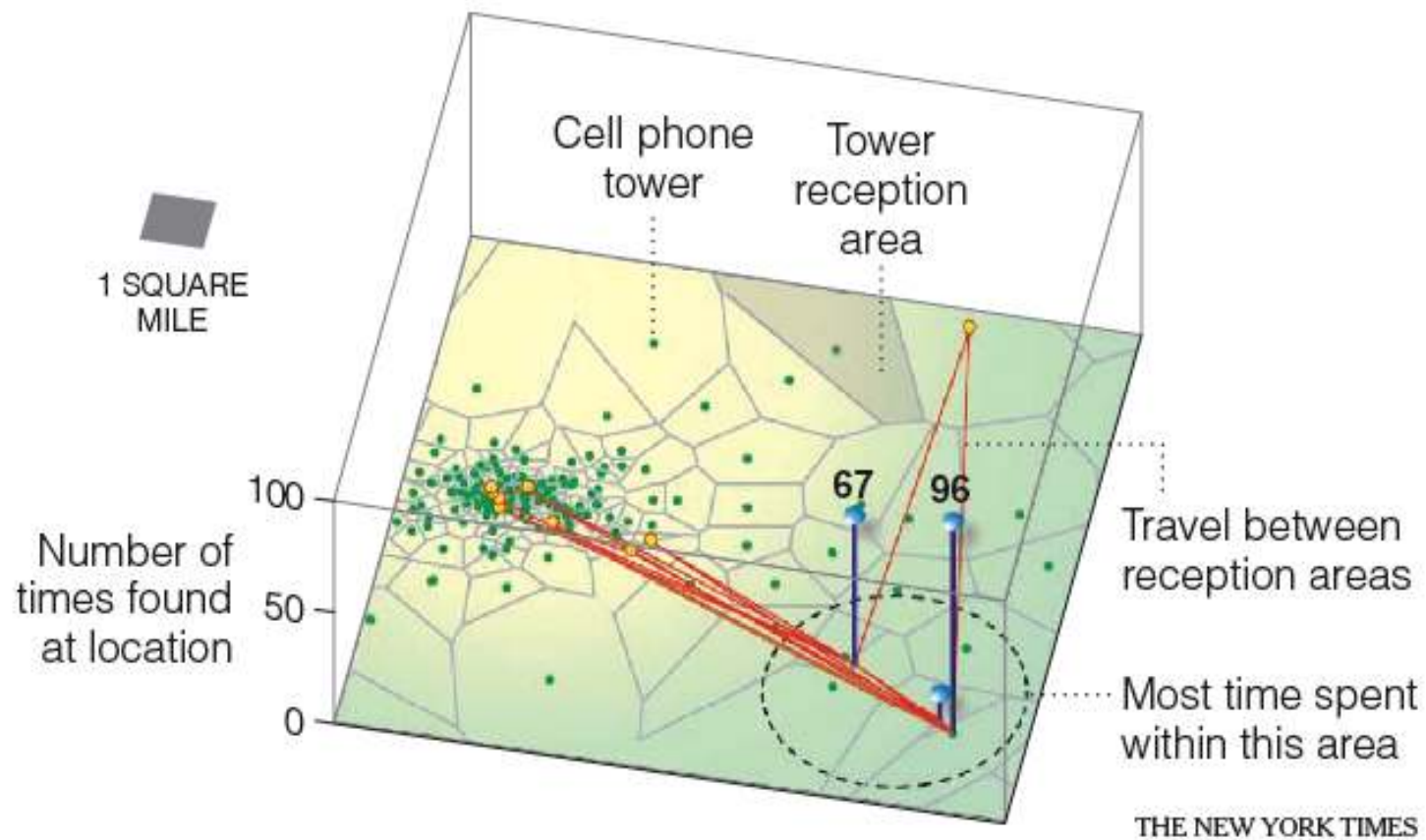
0%

Moto random

100%

Moto periodico

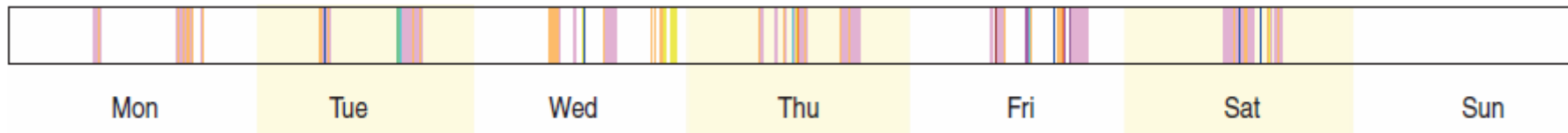
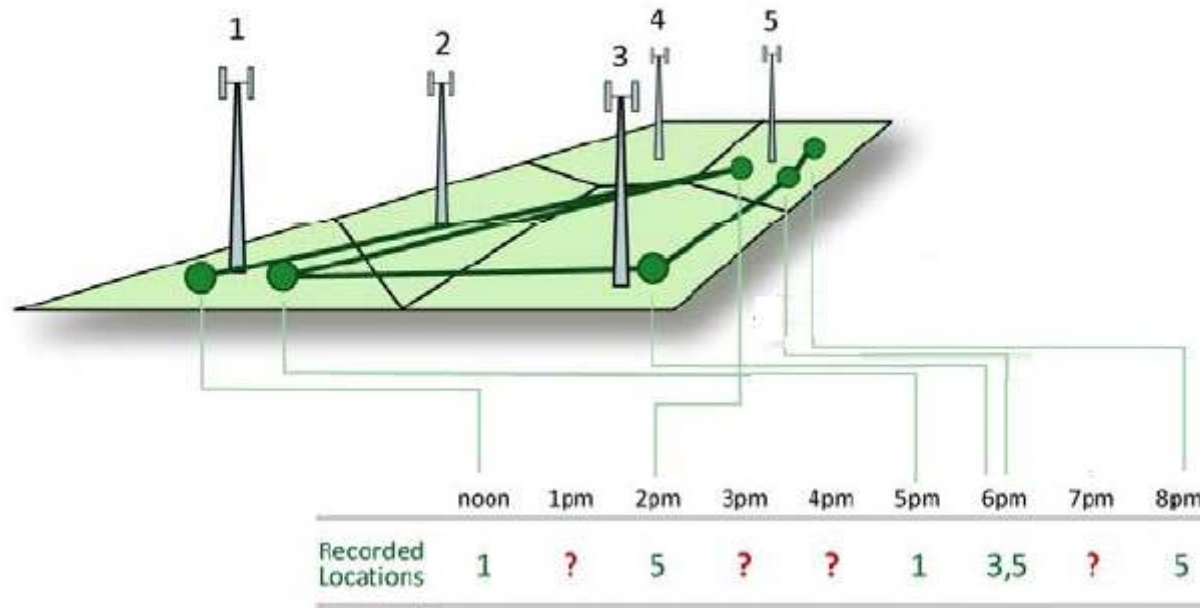




González, Hidalgo, Barabasi, *Nature* 2008

Entropia delle traiettorie umane

Recorded Trajectory



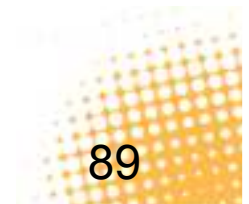
$$-\sum_{T'_i \in T_i} P(T'_i) \log_2 [P(T'_i)]$$



Il nostro tran-tran è molto prevedibile

■ **93%**

- Poca variabilità fra donne e uomini, ricchi e poveri, ...
- Song, Qu, Blumm, Barabasi, Science 327,108(2010)



privacy and ethical values in data mining



A new deal on data



**Personal Data:
The Emergence of a New Asset Class**

World Economic Forum, Davos, 2009

privacy and ethical values in social mining





Ethical values in social data mining

- **Privacy** – protection of human data to be mined
- **Fairness** – use of human (mined) knowledge
- **Trust, awareness and transparency** – use of data and knowledge



The new deal on data

- Quoting Alex (Sandy) Pentland (MIT) at WEF 2009
- **The first step toward open information markets is to give people ownership of their data.** The simplest approach to defining what it means to “own your own data” is to go back to Old English Common Law for the three basic tenets of ownership, which are the rights of
 - **possession,**
 - **use, and**
 - **disposal**



Tools for transparency and self-awareness

- **Full control of personal data / knowledge**

- From informed consent to awareness, support for the management of own personal data and knowledge

- **Data liberation**

- Right to withdraw personal data at any moment in full from any service provider

- **Oblivion**

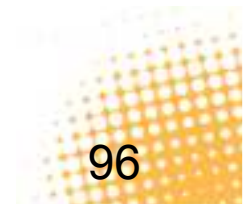
- Right to having personal data forgotten after a reasonable amount of time

- **Public good**

- Right to have full access to the collective knowledge

- **Identity management and protection**

- More knowledge, better shield against human hacking



Vizi privati & pubbliche virtù

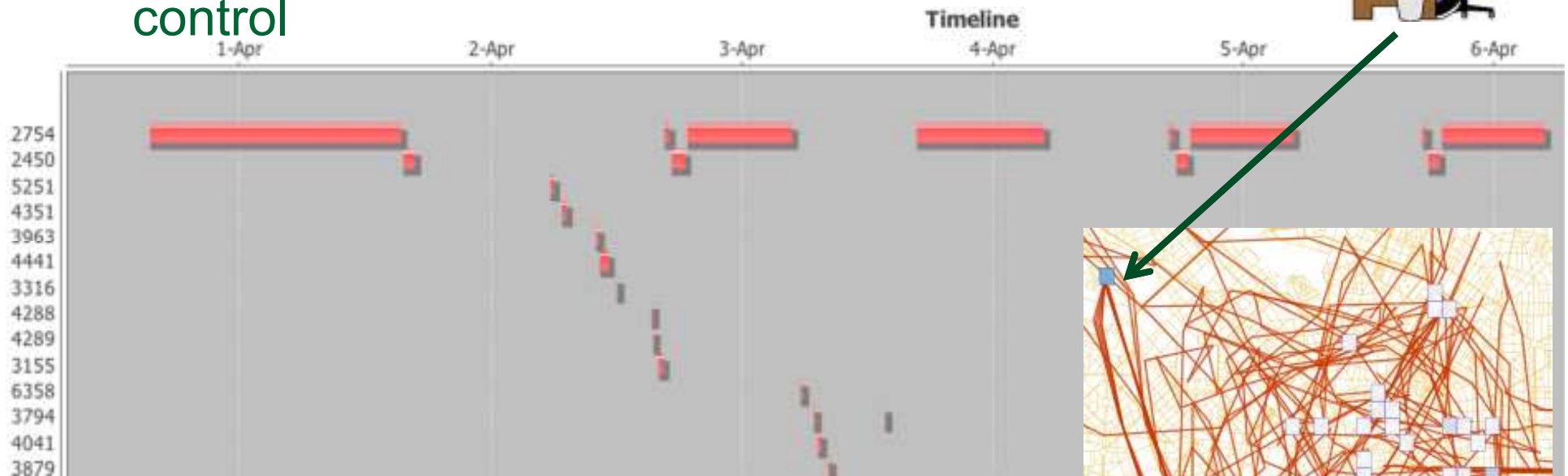
***Il dilemma fra privacy
e condivisione della conoscenza
al tempo di internet***



È possibile far coesistere il diritto alla protezione della propria sfera personale e il diritto alla conoscenza collettiva come bene comune?

Digital breadcrumbs and private sphere

- The digital traces that we leave behind may reveal personal traits that we do not want to share without control



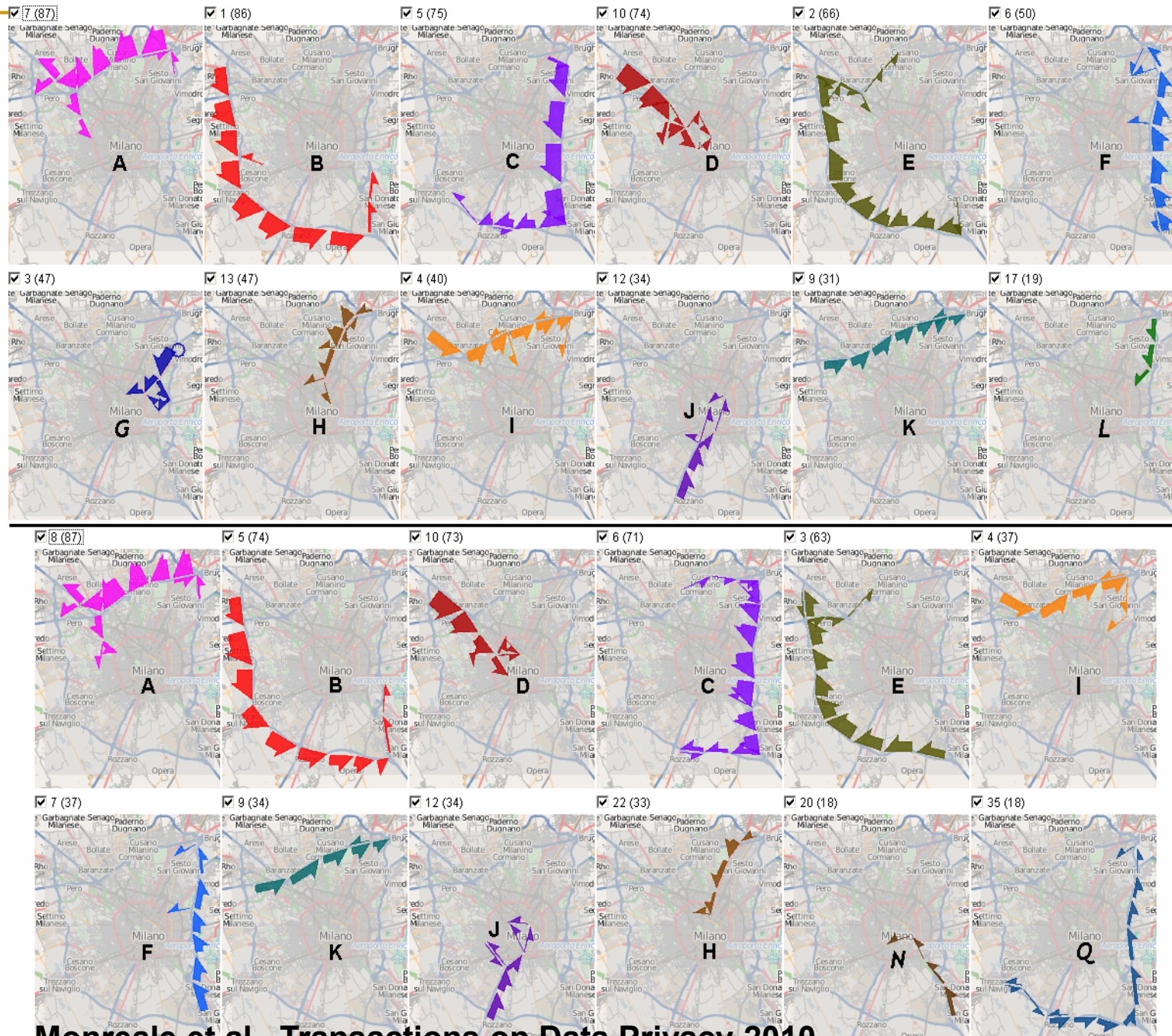
- Removing identities may not suffice to protect anonymity



Privacy by design principle

- In many cases (e.g., all previous questions!), it is possible to reconcile the dilemma between privacy protection and knowledge sharing
 - ❑ Make data **anonymous** with reference to social mining goals
 - ❑ **Use anonymous data to extract knowledge**
 - ❑ Only a little loss in data quality often earns a strong privacy protection



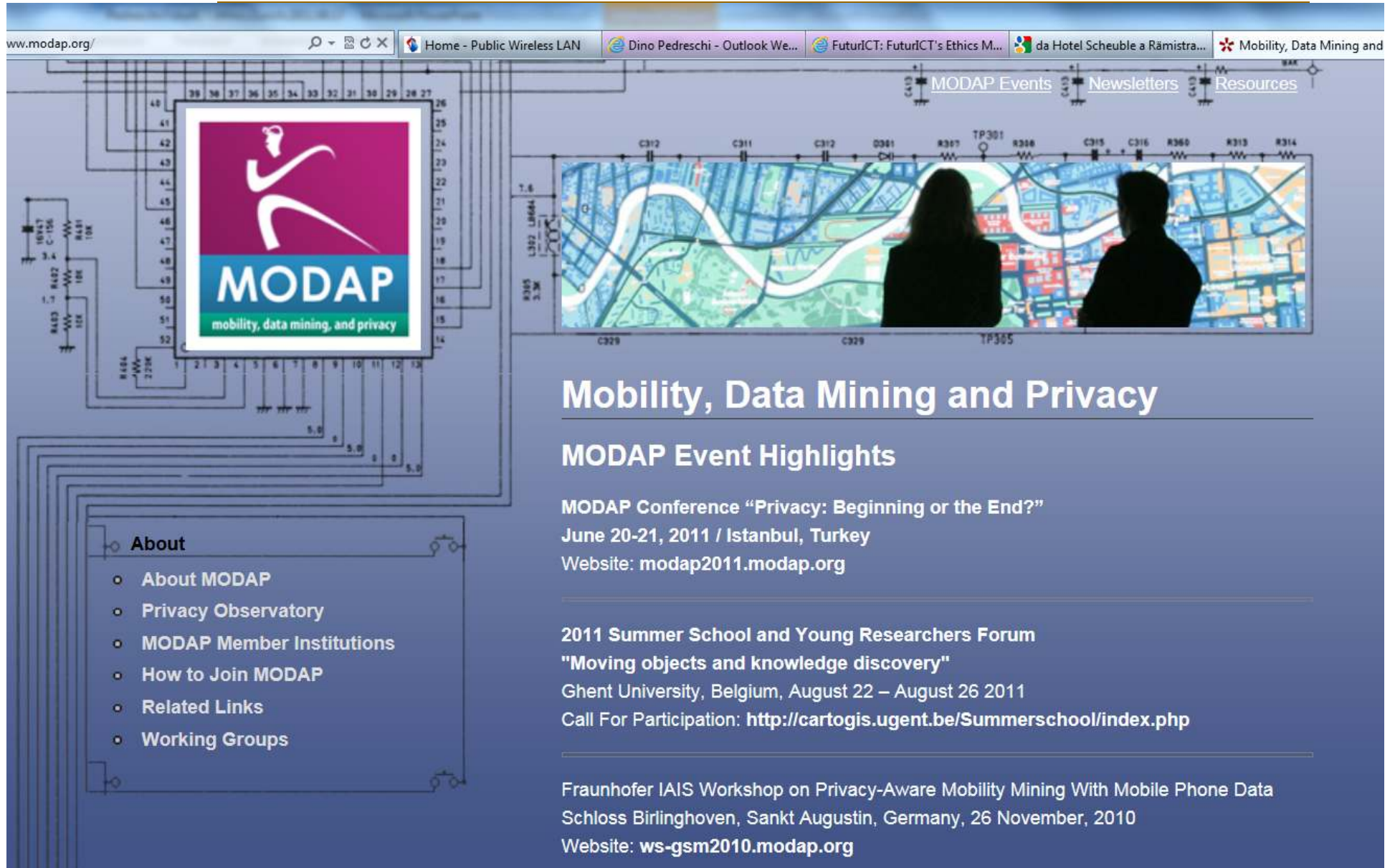


The reform of EC data protection directive

- New proposed directive submitted to European Parliament on Jan 25, 2012, approval process expected to complete within 2 years
- http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm
- Topics related the new deal on data:
 - Data portability
 - Right to oblivion
 - Profiling and automated decision making
 - Privacy-by-design



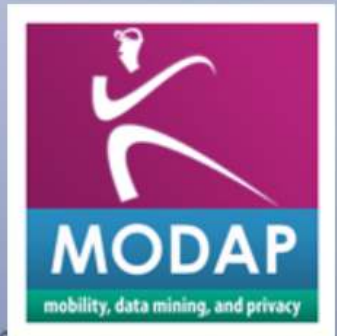
modap.org FET-Open C.A. – please join!



www.modap.org/

Home - Public Wireless LAN | Dino Pedreschi - Outlook We... | FuturiCT: FuturiCT's Ethics M... | da Hotel Scheuble a Rämistra... | Mobility, Data Mining and

MODAP Events | Newsletters | Resources



Mobility, Data Mining and Privacy

MODAP Event Highlights

MODAP Conference "Privacy: Beginning or the End?"
June 20-21, 2011 / Istanbul, Turkey
Website: modap2011.modap.org

2011 Summer School and Young Researchers Forum
"Moving objects and knowledge discovery"
Ghent University, Belgium, August 22 – August 26 2011
Call For Participation: <http://cartogis.ugent.be/Summerschool/index.php>

Fraunhofer IAIS Workshop on Privacy-Aware Mobility Mining With Mobile Phone Data
Schloss Birlinghoven, Sankt Augustin, Germany, 26 November, 2010
Website: ws-gsm2010.modap.org

About

- About MODAP
- Privacy Observatory
- MODAP Member Institutions
- How to Join MODAP
- Related Links
- Working Groups

» [Home](#)

Welcome to Privacy Observatory Magazine

Privacy Observatory Magazine is an electronic journal developed under the European project [MODAP](#). This journal combines scientific and legal expertise with a focus on issues related to data privacy and data protection . » [read more](#)

Sign in to [Modap Social Network](#) for staying up to date with the news from the project!

» SECTIONS

- » [Technology](#) (2)
- » [Law & Policy](#) (1)
- » [Industry & Business](#) (1)
- » [Case Study](#) (0)

» TAGS

[business](#) [economic approaches to](#)
[data privacy](#) [gps](#) [guidelines](#)
[industry](#) [iphone](#) [map](#) [michele](#)
[coscia](#) [privacy market](#) [terms](#)
[vehicular communication](#)
[webmaster](#)

» CURRENT ISSUE



Editorial

Today, we are faced with unprecedented opportunities of sensing and storing detailed data of human activities at a society-wide scale, because most of our activities are mediated by the ICT's.

» [read more](#)

» SEARCH

Search...



☒ Local ☐ Web

» [Advanced Search](#)

» LATEST NEWS

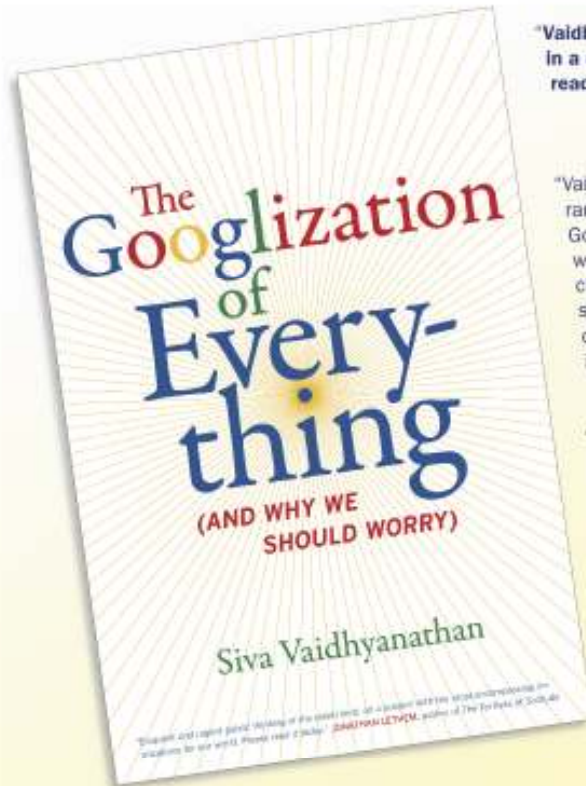
» [The Pros & Cons of Frictionless Sharing](#)

» [Dissecting Case 014 Exhibit B](#)

» [Lawmakers Seek FTC Pro of Facebook Post-Log C Tracking](#)

**"Finely written and engaging....
A book for anyone who has used Google."**

—Toby Miller, author of *Makeover Nation*



"Vaidhyanathan is everything you could want in a cultural critic: funny, fantastically readable, and insightful as hell."

—Cory Doctorow, author of *For the Win* and co-editor of *Bling Boing*

"Vaidhyanathan's lively, thoughtful, and wide-ranging book makes clear, in detail, how Google is reshaping the way we live and work. He finds much to admire, but also challenges us to not only use Google's services, but to go beyond them to create a new and genuinely democratic information order."

—Anthony Grafton, author of *Codex in Crisis*

"Thoughtfully examines the insiders influence of Google on our society.... As Vaidhyanathan points out, we must be cautious about embracing Google's mission and not accept uncritically that Google has our best interests in mind."

—Publishers Weekly, Starred Review

"A critically important book because it's really about the Googlization of All of Us.... A brilliant meditation on technology, information, and consumer inertia, as well as an ambitious challenge to change how, where, why, and what we Google."

—Dahlia Lithwick, senior editor and writer, *Slate Magazine*



At bookstores or www.ucpress.edu/googlization



UNIVERSITY OF CALIFORNIA PRESS

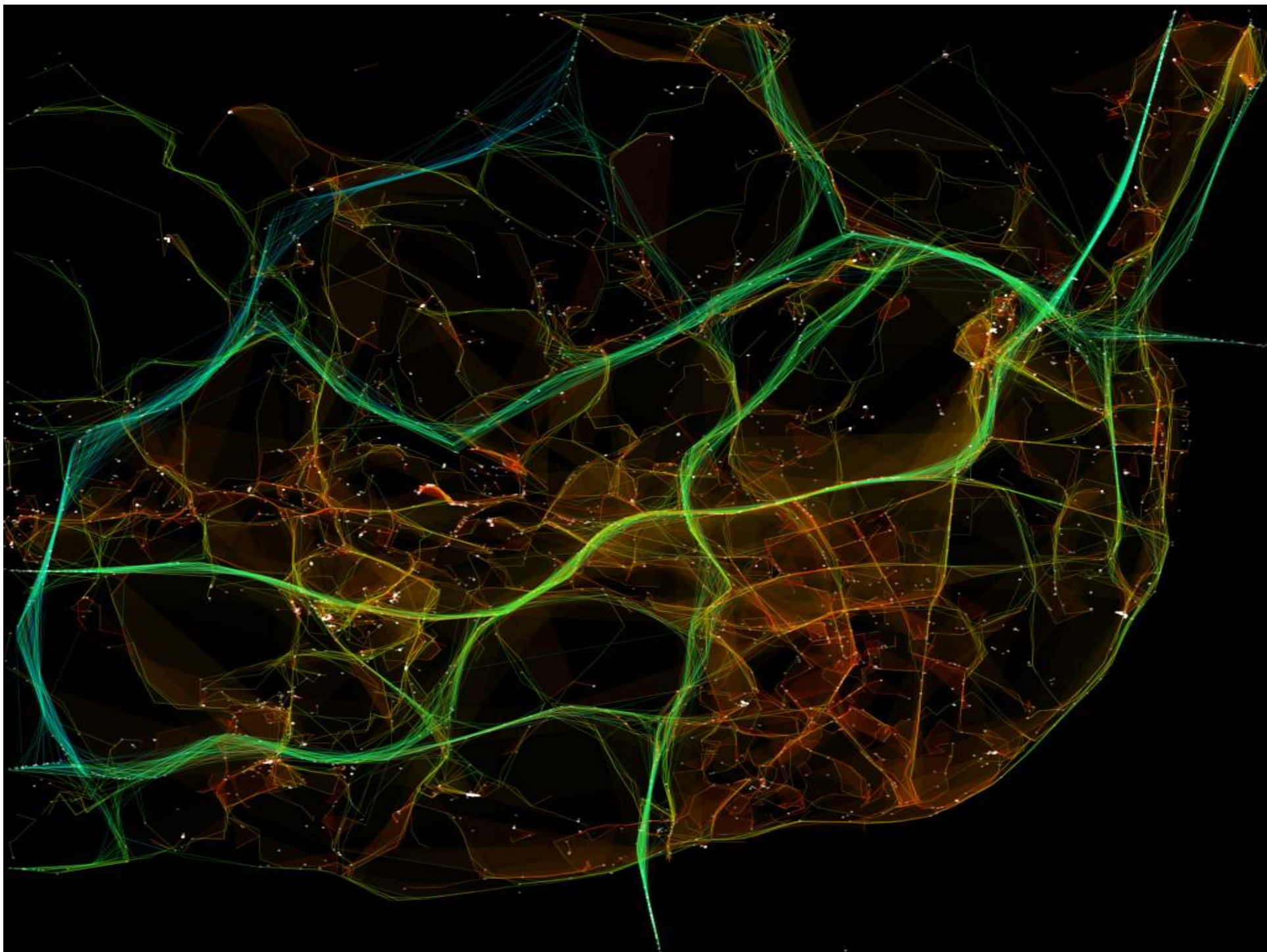
We are not Google's customers,
we are its **products**.

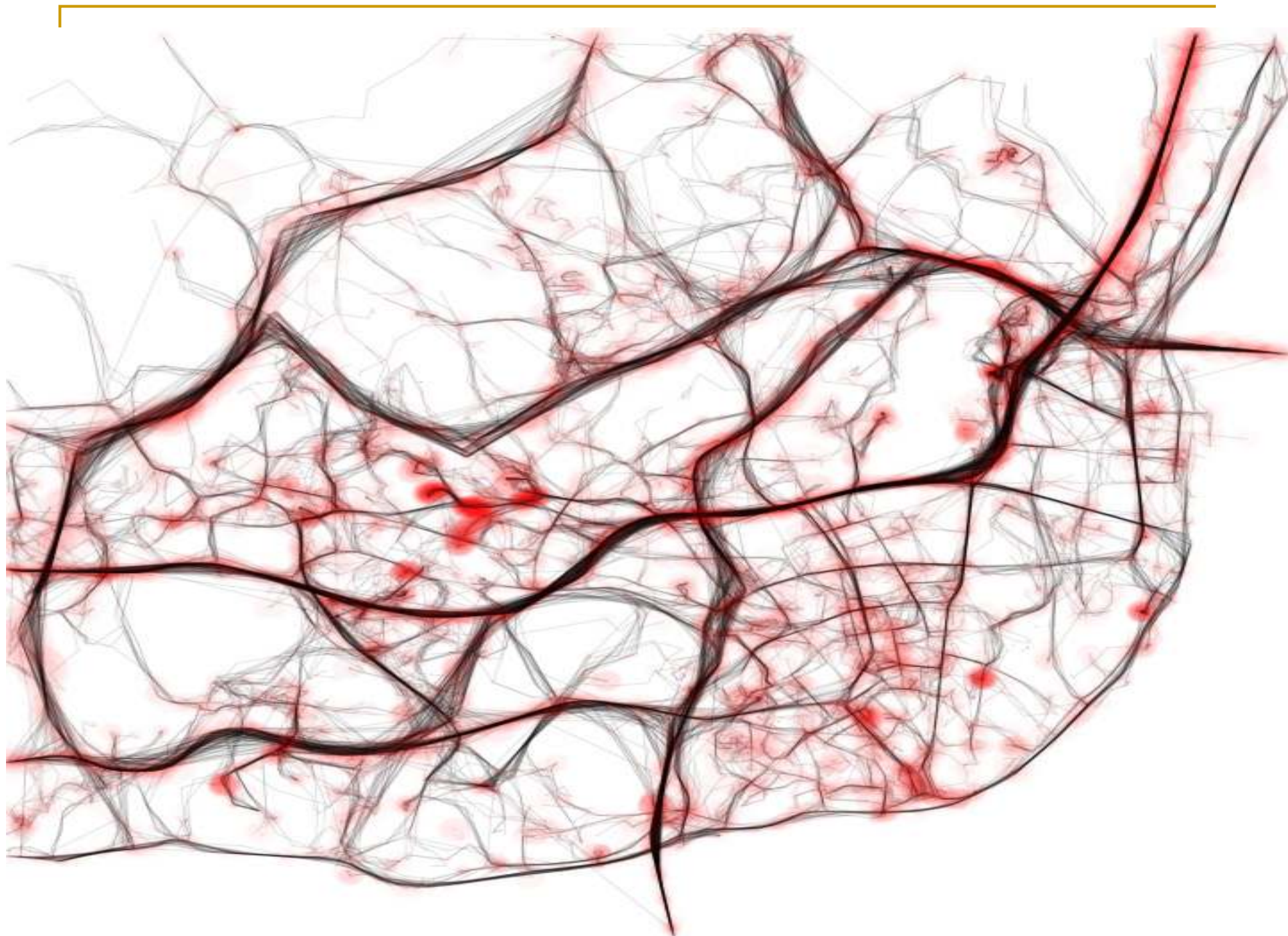
We – our fancies,
fetishes, predilections,
and preferences – are
what Google sells to
advertisers.

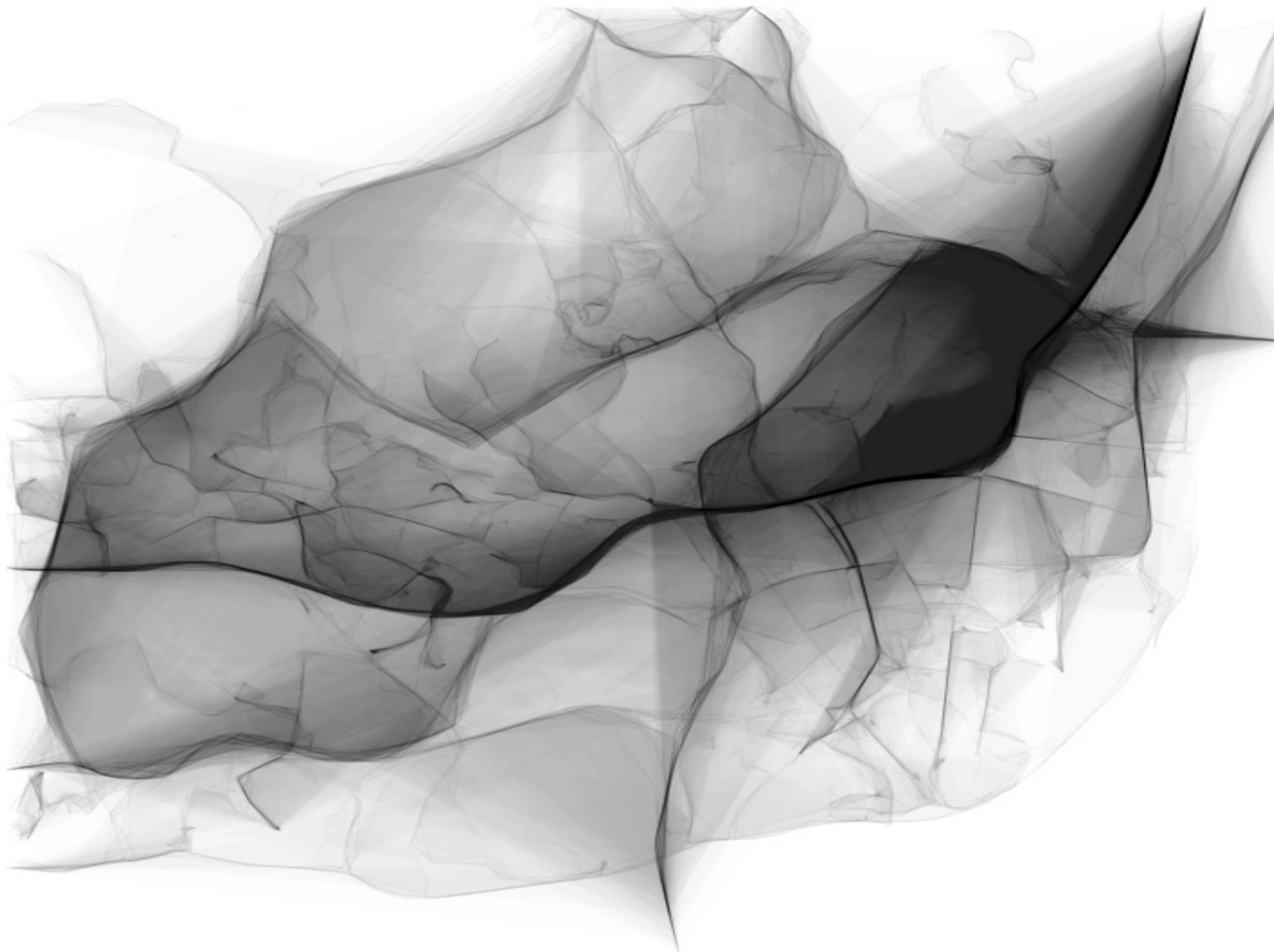
BIG DATA

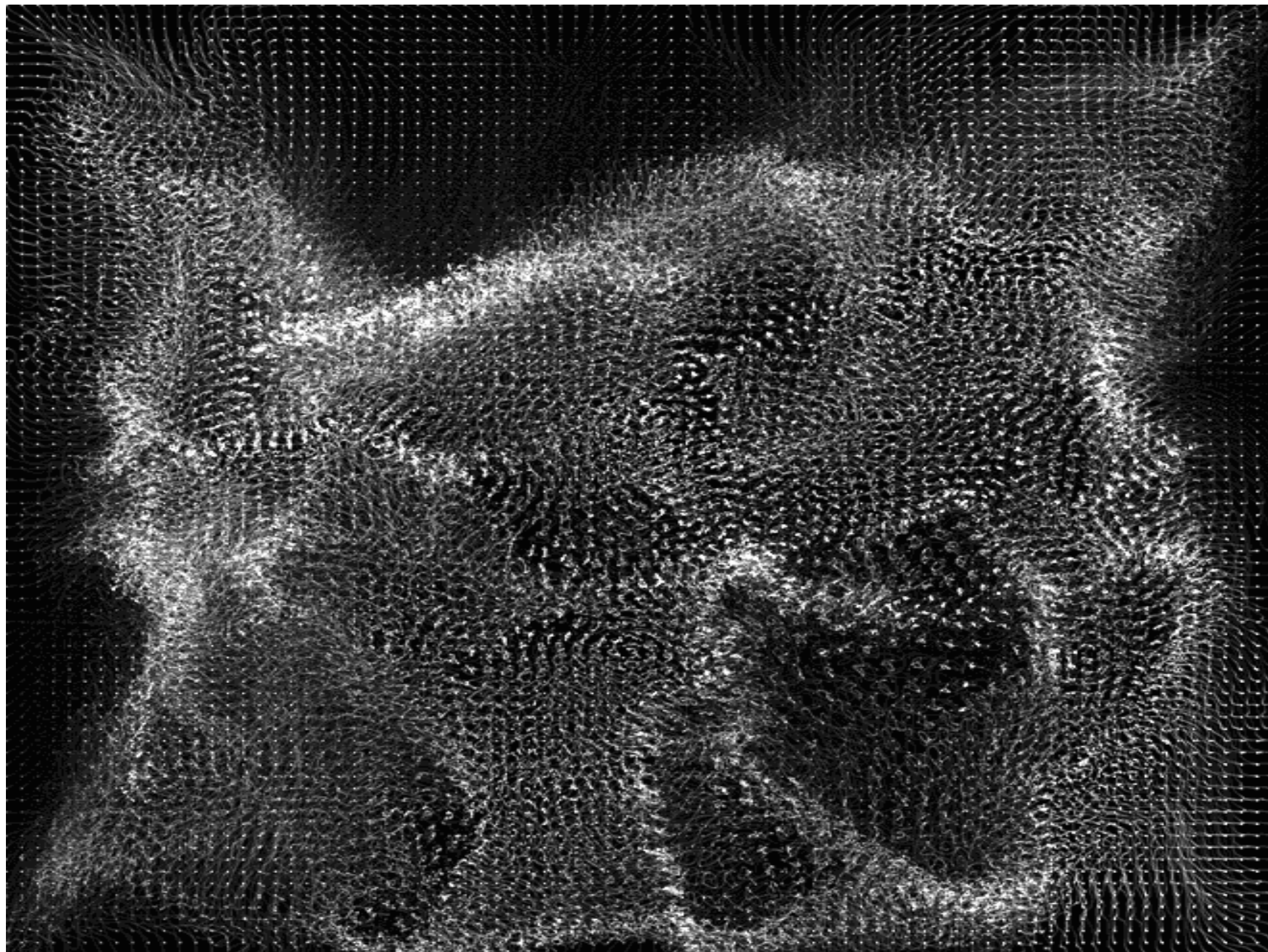
art











Pedro M. Cruz

- mondeguinho.com/master



Glitching Lisbon

Traffic jam

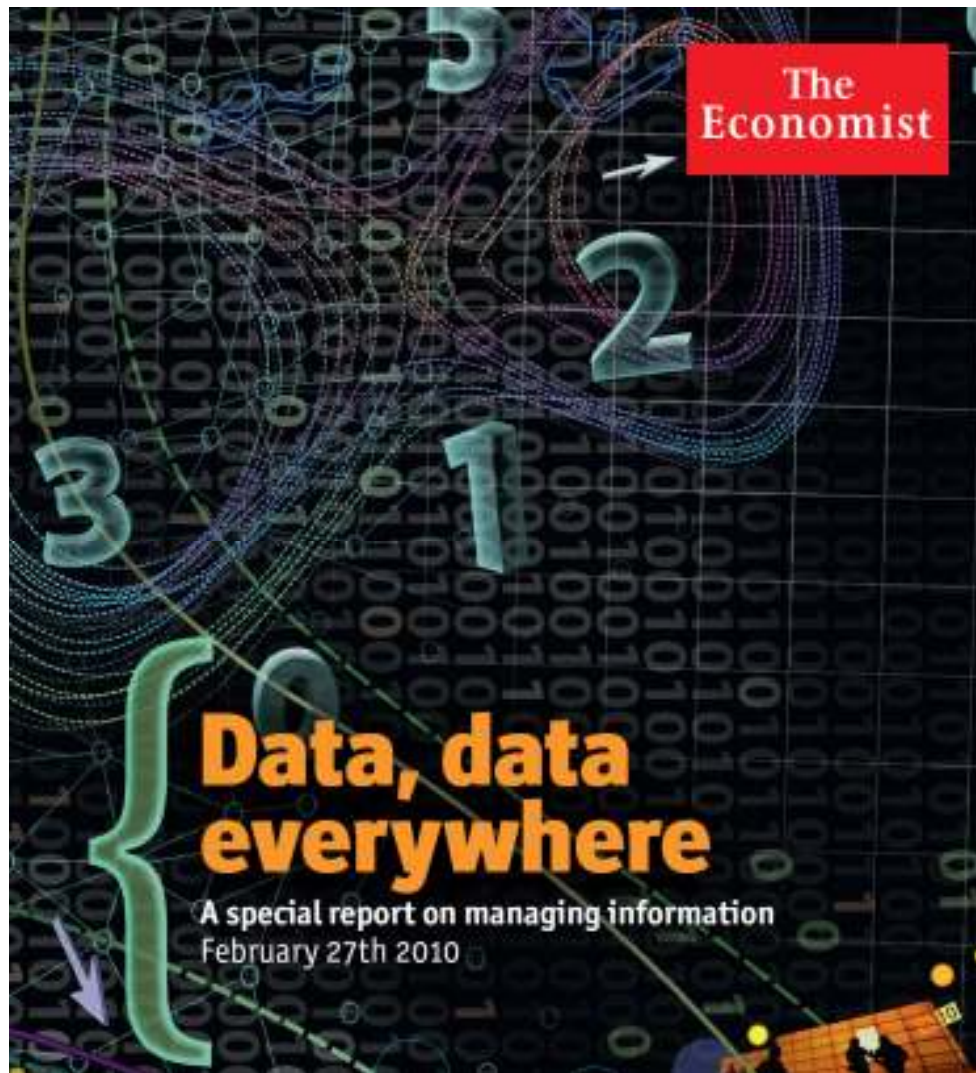
Estetica del traffico



BIG

DATA SCIENTISTS





- ... a new kind of professional has emerged, the **data scientist**, who combines the skills of **software programmer**, **statistician** and **storyteller/artist** to extract the nuggets of gold hidden under mountains of data.

- Hal Varian, Google's chief economist, predicts that the **job of statistician will become the "sexiest" around**. Data, he explains, are widely available; what is scarce is the ability to extract **wisdom** from them.



Think BIG!



19:00

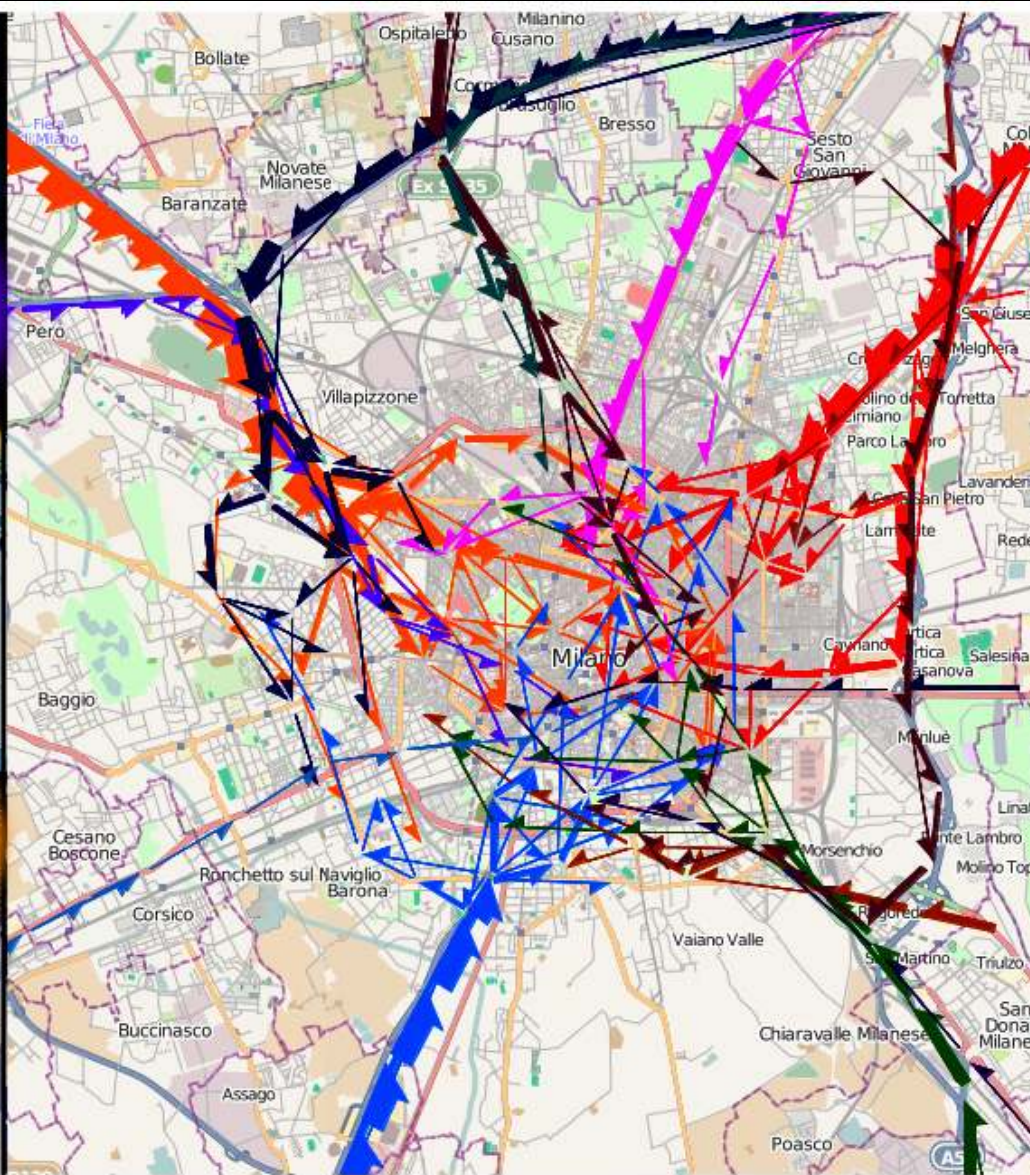
Nuit du nouvel an 2009 - SMS
Paris

urbanmiles



faberNovel

big data
anonymous
knowledge
footprints
- trajectories
data mining
privacy



kdd.isti.cnr.it