Big Data, Social Mining, Diversity, and Wellbeing

Big Data, Social Mining, Diversità, Benessere

Dino Pedreschi

Abstract Big data are the digital traces that we humans leave behind us continuously, as an aware or unaware side effect of the use of information and communication technologies for all our activities. By means of *social mining*, we have the chance of discovering knowledge hidden in Big Data, about social phenomena that hard to observe and measure. We discuss here how elusive and important social properties, such as diversity, collective intelligence and wellbeing, can be to some extent quantified and related to each other, thanks to novel analytics made possible by Big Data.

Abstract I Big Data sono le tracce digitali che lasciamo dietro di noi continuamente, come effetto collaterale, consapevole o meno, dell'uso delle tecnologie dell'informazione e della comunicazione per tutte le nostre attività. Attraverso il *social mining* abbiamo la possibilità di scoprire la conoscenza nascosta nei Big Data su fenomeni sociali difficili da osservare e misurare. Discutiamo in questo lavoro come alcune caratteristiche sociali sfuggenti e importanti, quali la diversità, l'intelligenza collettiva e il benessere, possono essere in qualche modo quantificate e correlate tra loro, grazie alla nuova analitica resa possibile dai Big Data.

Key words: Big Data, Social Mining, Diversity, Wisdom of the crowd, Collective Intelligence, Wellbeing, Social Network, Mobility

1 Big data analytics and social mining

Ours is a time of opportunity to observe and measure how our society intimately works: the Big Data originating from the digital breadcrumbs of human activities, sensed as a by-product of the ICT systems that we use everyday, promise to let us scrutinize the ground truth of individual and collective behavior at an unprecedented

Dino Pedreschi
KDD LAB, Dipartimento di Informatica, University of Pisa, Italy e-mail: pedre@di.unipi.it

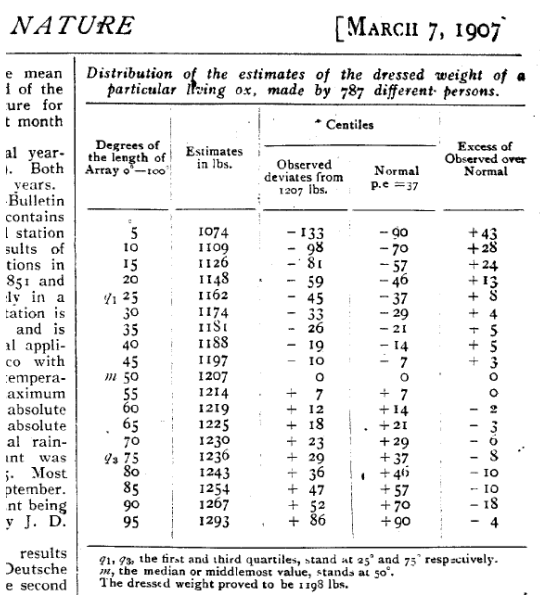
detail. Multiple dimensions of our social life have big data "proxies" nowadays. Our desires, needs, opinions and sentiments leave traces in the social media we participate in, in the query logs of the search engines we use, in the tweets we send and receive. Our relationships and social ties leave traces in the network of our phone or email contacts, in the friendship links of our favorite social networking site. Our shopping patterns and lifestyles leave traces in the transaction records of our purchases. Our movements leave traces in the records of our mobile phone calls, in the GPS tracks of our on-board navigation system.

Sensing big data at a societal scale, and the transparent interlinking of digital and physical reality, has the potential of providing a powerful social microscope, which can help us understand many complex and hidden socio-economic phenomena. It is clear that such challenge requires high-level analytics, modeling and reasoning across all the social dimensions above, an activity that it is increasingly often referred to as *social mining* [1]. Social mining is about making sense of Big Data by extracting meaningful information from large, messy and noisy data (originally produced for other purposes than analysis); social mining is also about narrating stories about the knowledge discovered in data. In this short paper I want to argue how data can tell interesting stories about an important social characteristic: *diversity*, and its surprising relationships with collective intelligence and wellbeing.

Why is diversity important? We learned that bio-diversity is crucial to the health of natural ecosystems and for the balance, or wellbeing, of plant and animal species that inhabit them. But diversity is a key concept also for the social ecosystems, made up today of highly interconnected people immersed in a plurality of information and communication technologies. It makes sense, a lot of sense, to speak of social diversity, especially for big data scientists and quantitative sociologists. Let me explain why, by means of two "data stories".

2 Diversity and the wisdom of the crowd

The first story begins a long time ago and with very "small" data. 1907, only two years after Einstein's *annus mirabilis*, is when the idea of the *wisdom of the crowd* is born. At that time Francis Galton, an English versatile scientist, publishes an article on *Nature* entitled "Vox Populi" [2], where he examined the outcome of a weight-judging competition at a cattle fair held in Plymouth, England, consisting in guessing the weight of a big cow on display. Galton retrieved the data of all the 787 bets made by people who participated in the competition; not only experienced farmers or butchers, but all sorts of people, attracted by curiosity, by the possibility of winning a prize or simply by the intellectual pleasure of the challenge. Galton imagined that the competition outcome could be viewed, in retrospect, as an experiment in which the crowd of 787 participants votes to collectively estimate what is the weight of the cow. Something similar to an election, the bet of each participant standing for her or his casted vote. Galton considered the median weight



NATURE [MARCH 7, 1907

Distribution of the estimates of the dressed weight of a particular living ox, made by 787 different persons.

Degrees of the length of Array 0°—100°	Estimates in lbs.	Centiles: Observed deviates from 1207 lbs.	Centiles: Normal p.e. = 37	Excess of Observed over Normal
5	1074	−133	−90	+43
10	1109	−98	−70	+28
15	1126	−81	−57	+24
20	1148	−59	−46	+13
q_1 25	1162	−45	−37	+8
30	1174	−33	−29	+4
35	1181	−26	−21	+5
40	1188	−19	−14	+5
45	1197	−10	−7	+3
m 50	1207	0	0	0
55	1214	+7	+7	0
60	1219	+12	+14	−2
65	1225	+18	+21	−3
70	1230	+23	+29	−6
q_3 75	1236	+29	+37	−8
80	1243	+36	+46	−10
85	1254	+47	+57	−10
90	1267	+52	+70	−18
95	1293	+86	+90	−4

q_1, q_3, the first and third quartiles, stand at 25° and 75° respectively.
m, the median or middlemost value, stands at 50°.
The dressed weight proved to be 1198 lbs.

Figure 1. Distribution of bets in Galton's experiment.

in the ordered list of bets, i.e., the prediction that has greatest consensus among the voters, in the sense that any other lower or higher prediction would have a majority of voters against. Interestingly, the median value of 1207 pounds (about 550 kg., see Fig. 1) was extremely close to the true weight of the cow, an almost perfect estimate. But even more surprisingly, Galton noticed that considering only the bets of the "experts" and repeating the same experiment, the estimation through the new median value turned out worse, i.e., more distant from the actual weight of the cow than the overall median! Hence the suggestion: a crowd can exhibit a very effective collective behavior in answering difficult questions, but the diversity of opinions in the crowd is essential. An elite of well-informed experts may actually have a smaller collective intelligence than a diversified crowd with diverse voices, in which preconceptions and "single thoughts" are not the rule. Since 1907 until today, many other studies have confirmed the *wisdom of the crowd* effect in different fields, such as stock markets, political elections, and quiz shows, discussed by James Surowiecki in his book [4]; in all such cases the diversity of the crowd is the essential factor. Continuing with surprises, Dirk Helbing and his team in Zurich has recently shown through social simulations, i.e., role-playing games run on computers by masses of digital players, that the ability of a crowd to reach correct conclusions (i.e., guessing the correct weight of the cow) decreases dramatically if the crowd is exposed to social influences [3]. In the experiment run by Helbing and colleagues, subjects in the crowd could reconsider their response to factual questions after having received information of the responses of other subjects. Although groups are initially "wise," knowledge about estimates of others narrows the diversity of opinions to such an extent that it undermines the *wisdom of the crowd* effect in different ways. The "social influence effect" diminishes the diversity of the crowd without improvements of its collective error. The "range reduction effect" moves the position of the truth to peripheral regions of the range of estimates so that the crowd becomes less reliable in providing expertise for external observers. In Helbing's words, "examples of the revealed mechanism range from misled elites to the recent global financial crisis."

It's a shocking discovery that that even mild social influence can undermine the *wisdom of the crowd* effect. It's a spotlight on the responsibility of the media and information system, whose influence on the public opinion with unilateral messages may promote conformity and marginalize the diversity of opinions, with the side effect of downgrading the collective intelligence of our society. Potential threats to diversity actually come not only from the press and TV, but also from the web search engines, recommendation systems and personalized marketing systems, which tend to fit people into profiles and stereotypes. As well as the excessive inequality, which widens the gap between the few who have too much and the many have too little, and that greatly increases the ability of elites to impose their single thought, especially in economics. These are all good reasons why we should protect socio-diversity, if we want to protect the health of our societies. We should spread powerful antibodies in the social network, in the media of all kinds, in the markets, able to prevent excessive conformity. Because greater diversity means greater collective intelligence.

3 Diversity and wellbeing

The second story is more properly a story of big data and tells us that diversity also means prosperity. A first example is the experiment of Nathan Eagle and colleagues [5], based on a massive dataset of phone record data, which allow reconstructing a synthetic proxy of the entire social network of England using the "who calls whom" links as a proxy of social ties. The idea is to measure the social diversification of each user on the basis of his/her phone calls: a user that calls always the same, few persons has a low diversification, compared to a user who often calls a vast network of contacts. Accordingly, each user can be associated with a numerical index of social diversity, or social entropy, and we can then aggregate this value in a given geographical area, such as a municipality or a province, for instance computing the median or average value of the social diversity of the inhabitants of that territory. We thus obtain a single indicator for each area, which can be compared with other indicators obtained through socio-demographic surveys conducted by official statistics institutes. Eagle and colleagues compared the social diversity of the local communities of England with a wellbeing composite indicator produced periodically

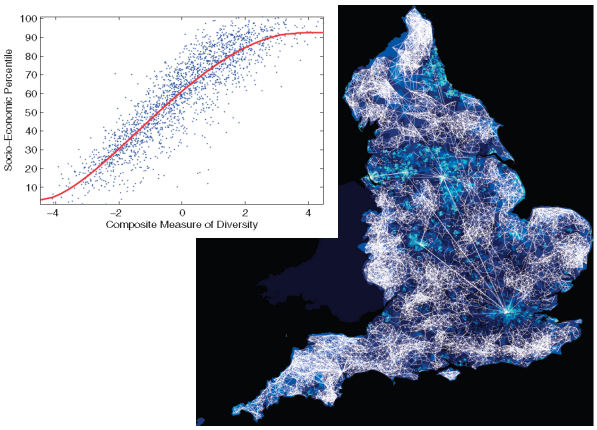


Figure 2. The correlation of diversity and wellbeing [5].

by the British government, which takes into account several socio-economic factors (income, education, health conditions, employment rate, etc.) The result, shown in Fig. 2, leaves no room for doubt: wellbeing is highly correlated to diversity; the two indicators grow (or fall) together.

Other, more recent studies from our research lab take into account mobility of people, again considering big data from either mobile phone call records or GPS trajectories from cars covering entire countries or large regions [6,7]. Our findings suggest an even stronger correlation between wellbeing and *diversity of movement* (compared to Eagle's): the greater the diversity of mobile behavior of the inhabitants of a territory, the greater its prosperity [8,9]. An explanation of this phenomenon may be that we move to perform our activities, and therefore an area exhibiting a large variety of motion is likely to be an area that offers many opportunities. In our studies, we analysed a nation-wide mobile phone dataset provided by Orange Telecom, covering entire France, which allowed a multidimensional study of human behaviour (see [8] for a preview of key results). Our purpose here is to understand the interconnection between the social network, the mobility patterns and the social and economic situation of French administrative districts, as measured by a deprivation index. Our analysis confirmed the existence of some already known patterns and revealed two new findings. First, the observation at district level of the three biggest cities in France (Paris, Marseilles and Lyon) uncovers a strong correlation between the social diversity of the personal contact network and the individual mobile diversity as shown in Fig. 2 (left). People who diversify their calls over the social contacts tend to have a more erratic mobility behavior.

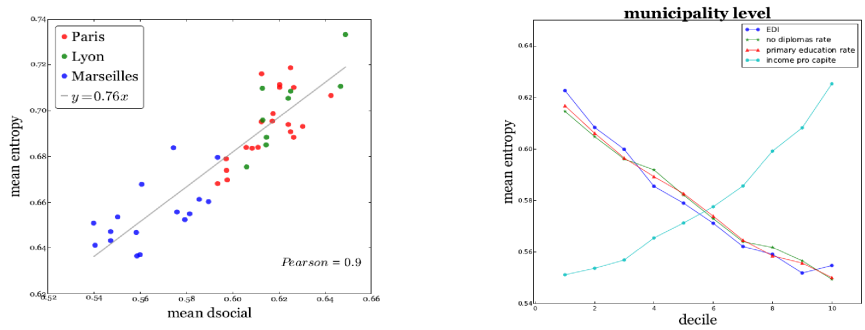


Figure 2: Correlation between mean social diversity and mean mobility entropy in main French cities' *arrondissements* (left); correlation between socio-economic indicators and mean mobility entropy in French municipalities (right) [8].

The second finding says that mobility diversity is strongly correlated with several indicators of the socio-economic level of geographic units, as shown in Fig 2 (right): a greater mobility entropy within a territory implies a lower value of various poverty indicators, and a larger per capita income. Our findings open an interesting perspective to study human behavior through mobile phone data, as new statistical indicators can be defined to describe and possibly predict (*nowcasting*) the

economic health of a territory. Our results are confirmed by another study [9] conducted on a different region (Tuscany) using different mobility data (GPS tracks from private cars).

4 The power of big data

The striking proliferation of Big Data, which characterizes our contemporary society, is paving the road towards a comprehensive picture of social behavior. Big data give us the opportunity to observe the network of social relations and movements, exposing the social fabric in which we are immersed. Sure enough, big data analytics and social mining have profound repercussions on many ethically sensitive issues, including privacy and data protection, data ownership and exploitation, transparency, self-awareness and self-empowerment, monopoly of knowledge of big corporations. In our globally interconnected world, we cannot afford to miss the opportunity disclosed by big data, we urgently need a techno-legal framework that allows for knowledge to become a safe, public good. Because Big Data may even allow us to capture elusive and fragile social properties, like diversity, and to fully understand their value.

References

1. Fosca Giannotti, Dino Pedreschi, Alex Pentland, Paul Lukowicz, Donald Kossmann, Jim Crowley, Dirk Helbing. A planetary nervous system for social mining and collective awareness. The European Physical Journal Special Topics 214(1) p. 49-75. Springer (2012). Available open access at http://link.springer.com/article/10.1140/epjst/e2012-01688-9
2. Francis Galton. Vox populi. *Nature* 75:7 (1907)
3. Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *PNAS* 108, 22 (2011)
4. James Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday Books, New York (2004).
5. Nathan Eagle, Michael Macy, Rob Claxton. Network Diversity and Economic Development. Science 328, 1029 (2010)
6. Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvo Rinzivillo, and Roberto Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The Very Large Data Bases Journal*, 20(5):695–719 (2011)
7. Fosca Giannotti, Dino Pedreschi. *Mobility, Data Mining and Privacy*. Springer (2008)
8. Luca Pappalardo, Giovanni Lima, Pierpaolo Paolini, Maarten Vanhoof, Zbigniew Smoreda, Dino Pedreschi, and Fosca Giannotti. Sociality, Mobility and Wealth indicators: a study in France based on big mobile phone data. In Proc. Mobile Tartu 2014, Mobile Data, Geography, LBS, Tartu Estonia July 2014 (to appear).
9. Lorenzo Gabrielli, Fosca Giannotti, Caterina Giusti, Stefano Marchetti, Luca Pappalardo, Dino Pedreschi, Monica Pratesi, Salvo Rinzivillo, Nicola Salvati. Small area model-based estimators using Big Data sources. Submitted for publication (2014).
10. (see also) Dino Pedreschi. La ricchezza della socio-diversità. Nòva – Il Sole 24 Ore, 7 febbraio 2014. http://nova.ilsole24ore.com/frontiere/la-ricchezza-della-socio-diversita
11. (see also) Dino Pedreschi. The treasure of socio-cultural diversity. TEDx Talks 2014 http://youtu.be/rBCeID5YT2E