

Statistics and Web: Who Needs Who?

Fabio Crescenzi, Daniele Frongia, Eric Sanna

Abstract Web usage has exploded over the last few years, and especially Social Web increased from tens of millions of users to many hundreds of millions of users. Great implications are expected in the incoming years especially on statistics. These implications will be surely bidirectional, from one side Web and Social Web will be demanding new statistical methods to monitor, sustain and enforce its growth, from the other side statistics will be demanding Web and Social Web to support crucial statistical activities. If the ability to visualize, share and interpret data may be clearly fruitfully increased, also data collection, monitoring and check of processes could take benefit of new web opportunities. In this paper we wish to investigate on these issues trying to give statisticians and IT researchers a list of substantive suggestions on fields which will be more probably demanding their actions.

Key words: complex networks, data sharing, social web, survey monitoring, survey management, web 2.0

1 Implications of Explosion of Web and Social Web on Statistics

What are the challenges for statistics in relation to the explosion of Web and Social Web today? Which research activities in the field of statistics will have more impact to sustain Web enforcement and conversely in which field of statistics Web can be most fruitfully used?

Google's chief economist Hal Varian in a recent interview makes the argument that with data in huge supply and statisticians in short supply, being a statistician has to be the "really sexy job for the 2010s". "While I would agree that the ability to visualize, interpret, and communicate data is clearly of great value, I still find that one of the biggest challenges organizations face today is how to aggregate the tremendous amount of scattered and inconsistent information out there. Before data crunching can begin, I

see a demand for people that can envision, build, and manage dynamic databases that can grow alongside changing business needs and evolving data sources” [16].

If explosion of Web and Social Web certainly requires new impetus to develop strategies aiming to face some crucial issues as the data validation and quality certification, and as the assessment of data confidentiality, it must be stressed that the new paradigm will not jeopardize the current quality standards: in fact, in a Social Web environment, producers of official statistics allow users to re-use data in other web sites or blog, but they still remain owner of those data. In other words, data sources remain the "official one", while each user might become producer of a data interpretation and/or visualization. It remains the risk of bias arising from media and users, which has to be monitored and managed through ad hoc initiatives (as in the traditional non web data case) adopting codes of practices and making mandatory, for users, the description of data processing.

In this paper we wish to investigate these issues trying to give statisticians and IT researchers a list of substantive suggestions on fields which will be more probably demanding their actions. We will propose at the meantime a possible classification of these actions in a concise taxonomy.

2 How Statistics May Help Web and Social Web?

In a nutshell, main fields in which statistics are useful to the Web concern measuring web usage and website performance. New impulsion is needed in the research of new application of graph theory to complex networks.

2.1 *Measuring Web Access and Usage*

Measurement of web access and usage are fields in which statistics may greatly help Social Web and above all have to be reported. The Web has changed fundamentally our life and it is vital for nowadays society. The design of future Web (and future society) strongly needs new methodological advances.

A new field of research in statistics has to be dedicated to the measurement of web traffic and usage indicators, to propose new model to produce good forecasts of the number of people who have web access and the number of people who may be assumed to be as web users. Online commerce, but also all business sectors that may benefit from the development of online services, need to know the future evolution of the use of the Web in its particular segments. The study of websites birth and death could give rise to a further field of research, to investigate on “websites demography”, but this requires a more precise definition of the objects and the events involved (e.g. which is the time to assume as dead time?).

2.2 *Defining and Analyzing Website Ecosystem*

Web analytics is the collection, measurement, reporting and analysis of information about users habits in order to understand and optimize website usage. On-site analytics refers specifically to the web based measurement, analysis and reporting of specific data on a given website.

The Web is no longer just a collection of individual sites but it can indeed be seen as a set of "online ecosystems", each of one consisting of a set of thematic web sites: in this case a new challenge is to determine new key performance indicators to evaluate trends of sites and blogs within such ecosystems. Specifically, new methods aim to measure positioning considering infrastructural features of websites [7].

After considering the usual indicators such as user traffic, today we are able to analyze also the "position" of a site in the network of the WWW. A web site, in fact, could be well "positioned" in the network, that is, its "centrality" as the node could be very high. To determine such position we need to consider a new set of indicators such as Google PageRank, Twitter, Wikipedia external links, etc. Alternatively, there is a new generation of tools that measure and monitor the position of brands, blogs or web sites with an approach based on both Googledatabase and traditional indicators of social network analysis.

Statisticians are expected to contribute greatly in producing more credible indicators in this field and new measurement methodologies [10,11].

2.3 *New Application of Graph Theory on Complex Networks*

During the last decades, the network analysis has received a great boost by mathematicians and physicists and, at the end of the '90, major studies were conducted to determine the structure of complex networks [3,4,6,8].

Today, the contribution of statisticians to develop certain themes (such as investigation on the structure of the Web [14], new models to distinguish between essential and redundant information, resistance of complex networks [1], estimation of the size of the Web and the number of pages and sites [12, 17] etc.) could highly increase possibility to better know the present and the future of the World Wide Web.

A suggestion for a first action is to strongly enforce the efforts of statisticians to produce new indicators for monitoring the state of the web, including new methods for longitudinal analysis. The objective is to produce continuously and in a certified way a minimum set of key indicators to facilitate the statistical study of the web in its evolutionary trends.

2.4 *Data Integration and Semantic Web*

A fundamental aspect for social and wide use of statistics is availability of certified data sources, which can be also interrogated without any human intervention. This can be made by appropriate endpoints, through specialized software and query languages which enable users and systems to interrogate large data repositories. These endpoints,

providing a machine-to-machine interaction between systems, allows to extract subsets of requested data and encapsulate it in other applications, websites or blogs, stimulating the creation of new user-generated contents. Data remains on servers of data producers (such as a national institutes of statistics) and users can embed data and application (e.g. widget) into other sites or blog. When data change, all applications update automatically.

The availability of these services determines, for social applications, the possibility to integrate, in the same layout, data from different sources and produced by different organizations. It also allows to collect, explore, compare and geocode data, integrating them with maps, making the results immediately accessible via Internet.

Statisticians will probably be requested to collaborate also on Semantic Web, an evolution of the World Wide Web focused on relation to any Internet resource to each other, and to Linked Data, a set of technologies to achieve this objective for data, to create a Web of data [5].

3 How Social Web May Help Statistics?

On the converse, it is not difficult to foresee that, in the incoming years, new web applications will produce a genuine revolution in statistics. If the new potential use of Web is intuitive in dissemination and in statistical data sharing, areas not sufficiently explored concern its use in many of the major stages of statistical data production process.

A further issue is here only mentioned and concerns the communication of statistics. In last years it has been widely discussed a paradox which concerns statistics: a poor public image coexists with a huge social need of statistics and its “ubiquity” [9]. Web, for its characteristics, really pays to be used for the purpose of inform on good results achieved in application of statistics. To this purpose could be used, ad example, social networking platforms oriented to statistics.

3.1 *Sharing, Integration, Treatment, Analysis and Visualisation of Statistical Data*

New sites have been launched to allow all web users to upload and integrate data, share and visualize them, and talk about their use with other people. Among them IBM Many Eyes and Swivel, online services already used by some international governmental organizations.

The development of web specialized environments, able to integrate statistical information at various levels of structure, from free text data-formatted to rigidly structured data, are crucial to answer to the biggest challenges announced by the Google Chief economist Hal Varian: how to face the needs of aggregation of the tremendous amount of the scattered and inconsistent information existing [16].

One of latest response to these challenges are online statistical-oriented services and business intelligence tools (such as Good Data [13]), which enables to build and manage a multi-dimensional data model from different data sources, providing a

collaborative environment and tools to analyze data, realize personalized output tables and graphs and share it.

The growth of the Web require new impetus to develop strategies for the certification and the validation of data, taking into account the most important dimensions of the quality as accessibility, comparability of integrated data, etc.

Further, will increase the need of new methods and organizational solutions to reduce the disclosure risk of data shared online.

3.2 *Supporting Statistical Surveys Activities*

Online surveys are today widely used in data collection and survey methodology deals web surveys methods more and more. The increasing spread of Web opens the door to future complete online and paperless surveys, but in the short term pushes toward a larger use of multi channel surveys: in fact to reach the part of respondents not having web access, the web collection, enlarged as soon as possible, will have to coexist with a still significant portion of non web data collection. Survey management systems will have to increasingly integrate control functions of the entire production process and web applications may help to hold together and control the inputs of different data collection channels and these may help greatly in monitoring and increasing response rates [15].

In future, beside traditional Web, also Social Web could play a role in data treatment and management of statistical surveys: 1) to convince and to push people to participate to statistical surveys helping to reduce non response rate; 2) to capture information on respondents to investigate on the goodness of data and to be used in the data treatment; 3) to share real time information and signals to help surveys monitoring and back-office activities.

Survey methodologist will be asked to inquire on new methods to catch from social web information useful for survey processes and new methods to employ these information to improve data quality. Some of these methods could be integrated in future generations of online statistical surveys monitoring and management systems.

4 Conclusions

In the incoming years it's expected a great joint effort of statisticians and IT researchers at least in these fields: 1) to realize the social evolution of statistics 2) to develop methods and tools to interrogate and analyze large data sets; 3) to analyze and systematize theories and concepts about complex networks; 4) to develop web based statistical tools and platforms.

At the same time social web applications will be able to capture and store, in secure way, large amount of data about characteristics and habits of users which may be fruitfully used in the survey process.

The idea to "throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science can not" [2] found several supporters, but it's destined to remain an utopia without the availability and use of appropriate statistical methods and tools for strict quality check of data.

References

1. Albert R., Jeong H., Barabási A. L.: Error and attack tolerance of complex networks, *Nature* 406, 378-382 (2000)
2. Anderson C.: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired* (2008)
3. Barabási A. L., Albert R.: Emergence of scaling in random networks, *Science* 286, 509 (1999)
4. Barabási, A.L.: La scienza delle reti, Einaudi, Torino (2004)
5. Berners-Lee T.: Semantic Web and Linked Data, speech at TED Conference (2009)
6. Bianconi G., Barabási A.-L.: Bose-Einstein condensation in complex networks, *Phys.Rev.Lett.* 86:5632-5635 (2001)
7. Frongia, D.: Il successo di un sito web? Non solo una questione di traffico: i casi Istat, SIS e Sistan, *Sis Magazine* (2010)
8. Frongia, D.: Intervista a Ginestra Bianconi sulle reti complesse, *SegnalazionIT* (2008)
9. Hand, D.J.: Modern statistics: the myth and the magic, *J. R. Statist. Soc. A* (2009)
10. Koizumi D., Matsushima T., Hirasawa S.: Bayesian Forecasting of WWW Traffic on the Time Varying Poisson Model, in *Proceeding of The 2009 International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA* (2009)
11. Polanco X., Roche I., Besagni D., Web Usage Analysis: New Science Indicators and Co-usage, *Seminaire VSST 2006, Lille, France* (2006)
12. Reka A., Hawoong J., Barabasi A.L.: Diameter of the World-Wide Web, *Nature* (1999)
13. Sanna E.: Business Intelligence con Good Data, *SegnalazionIT* (2009)
14. Serrano M.A., Maguitman A., Boguna M., Fortunato S., Vespignani A.: Decoding the structure of the WWW: facts versus sampling biases, *ACM Transactions on the Web (TWEB)* 1, 10 (2007)
15. Sindoni G.: An online system for multi-channel, register-based census data collection (2009), *Conference of European Statisticians, Group of Experts on Population and Housing Censuses, Twelfth Meeting, Geneva* (2009)
16. Varian H.: On how the Web challenges managers, *Business Technology Office* (2009)
17. VV.AA.: World Wide Web, Statistics section, *Wikipedia* (2010)