Round table *"Data in complex systems research. Problems and opportunities in biological and social systems"*. Palermo, Tuesday April 8th 2008.

The talks presented at the Conference raised several points that are briefly summarized hereafter:

1) Does the necessity of performing research with proprietary data have an impact on the reproducibility of scientific results within an academic community? Is this a specific aspect related to complex systems research? Or, rather, this is just an aspect of a typical problem researchers face, for example, in industrial research.

2) Concerning privacy issues are there crucial differences between the management of records dealing with biological and social information? Which kind of data is today better protected and nevertheless accessible for academic research?

3) Is it possible to use a set of real data to build up a synthetic population that is representative of the statistical properties of the real population and can be freely distributed among researchers without any privacy problem? Is the above scenario a realistic one for research in many complex systems?

Starting from these questions the discussion has focused on these aspects and on other related problems. Several participants expressed their views and hereafter I am attempting to synthesize some key aspects of the contribution of some of them.

Yakovenko: The proprietary nature and related privacy issues of many interesting datasets have a huge impact on complex systems research. The tools proposed in computer science to perform research without a full control of the analyzed data (see the Marchiori's talk) are certainly encouraging.

Shavitt: It should be kept in mind that with suitable data processing and data mining of anonymized data the discovery of information whose privacy is protected could reach a percentage as high as 87% of the considered data set.

Kertesz: Concerning the difference between biomedical data and social data my impression is that they are totally different and should be approached in a different way. One way to solve the problems of privacy and proprietary nature of data can be to prepare and share artificial synthetic data. Another aspect the research community should focus on is the standardization of similar data. Currently there is no attempt for standardization of social data whereas, in some case, a certain degree of standardization is present in biomedical data.

Monti: There are intrinsic differences between biomedical data and social data. Genomic data are more costly than many social data and the dimensionality of these data are also quite different.

Muller: Also financial data can be quite expensive and huge efforts are often devoted to achieve the best result in a specific area of financial interest.

Eubank: A problem to be considered concerning the access to proprietary for research purposes is that sometimes the access to data is restricted to just the investigation of a quite specific problem that is usually proposed by the owner of the database. When a different research is considered a new negotiation need to be done with the owner of the database.

Onnela: Concerning the problem of scientific reproducibility of results obtained by using a proprietary (or restricted access) database there is no easy solution. One possibility might be to accept a weak form of reproducibility to be expected in the long run and perhaps obtained by considering investigations performed in different (although analogous) databases.

Shavitt: I am quite skeptical about the possibility that different datasets might allow a certain degree of reproducibility of scientific results in the long run.

Kertesz: A very useful approach could be to select, whenever possible, publicly accessible data for benchmarking of different empirical procedures and models.

Yakovenko: It should be pursued an editorial policy from scientific journals promoting the sharing of data used in published papers. Whenever possible the data, or at least part of them, should be accessible and downloadable as it is the case for several biological and biomedical researches.

Eubank: In NSF grants you need to make data public after 2 years of the grant.

Onnela: The problem of reproducibility of results is certainly not specific to complex systems research. Similar problems of reproducibility are also encountered in computational physics.

Lancet: Data in biology and social systems share points of similarity especially from the side of system biology where there is a need for pre-biological simulations and artificial data sets to be investigated.

Weisbuch: A repository of publicly accessible data in complex system research would be helpful.

Shavitt: Useful links to publicly accessible database already exist. Indeed you just need Google.

Yakovenko: Indeed our community should promote the development of a repository for data.

Liljeros: Continuous and qualified contribution to a repository is time and effort demanding. Why one member of the research community should contribute to the repository? Which should be the rational motivation for that?

Lancet: To achieve something successful one would need a centralized institution. Funding agencies should consider the potential usefulness of a similar project.

Mantegna: A data repository on complex systems research should face the multiple nature of data of interest in different areas of research.

Bazzani: Sometimes a company is not full aware of the value of their data and while using the data the perception of the company about their value changes due to the fact that the research performed reveals such value.

Kertesz: There are various data treated differently with respect to the disclosure policy of the proprietary of databases. For example for most of the financial data one can negotiate the access and use of the data. For mobile phone data it is completely different. The company doesn't allow making any disclosure.

Eubank: There are sets of data researchers cannot use for any research. Also some data originally publicly available can after time be very difficult to access. An example concerns the data about infrastructure, which disappeared from the Internet after September 11.

Mueller: In finance most of the relevant data are proprietary data. For example in the insurance industry almost all the relevant data are proprietary data.

In summary in the discussion emerged a series of problems related with the approach to data in complex systems research. Data in different research areas of complex systems present both similar problems of access, need of standardization, creation of sets for benchmarking, promotion of the dissemination of data and also other issues that are typically quite specific of the research area (for example insurance, mobile phone traffic, finance, genomic or infrastructure). Perhaps an effort attempted by a major funding agency centralizing an experimental repository on data in complex systems could over the years tackle common problems of access, contracting and standardization and, in parallel, consider specific aspects of each major research area.