|  |
| --- |
| %% USPSC-Introducao.tex |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| % ---------------------------------------------------------- |
|  |

|  |
| --- |
| % Introdução (exemplo de capítulo sem numeração, mas presente no Sumário) |
|  |

|  |
| --- |
| % ---------------------------------------------------------- |
|  |

|  |
| --- |
| % \chapter[Introdução]{Introdução} |
|  |

|  |
| --- |
| \chapter{Introduction}\label{ch:int} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The first studies dealing explicitly with human interaction networks |
|  |

|  |
| --- |
| date from the nineteenth century while the foundation of |
|  |

|  |
| --- |
| social network analysis is generally attributed to the psychiatrist Jacob Moreno in mid twentieth century~\cite{moreno,newmanBook}. |
|  |

|  |
| --- |
| With the increasing availability of data related to human interactions, research about these networks has grown continuously. |
|  |

|  |
| --- |
| Contributions can now be found in a variety of fields, from social sciences and humanities~\cite{latour2013} to computer science~\cite{bird} and physics~\cite{barabasiHumanDyn,newmanFriendship}, given the multidisciplinary nature of the topic. |
|  |

|  |
| --- |
| One of the approaches from an exact science perspective is to represent interaction networks as complex networks~\cite{barabasiHumanDyn,newmanFriendship}, with which |
|  |

|  |
| --- |
| several features of human interaction have been revealed. |
|  |

|  |
| --- |
| For example, the topology of human interaction networks exhibits a scale-free outline, |
|  |

|  |
| --- |
| which points to the existence of a small number of highly connected hubs and a large number of poorly connected nodes. |
|  |

|  |
| --- |
| The dynamics of complex networks representing human interaction has also been addressed~\cite{barabasiEvo,newmanEvolving}, but only to a limited extent, since research is normally focused on a particular metric or task, such as accessibility or community detection~\cite{access,newmanModularity}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| There are numerous articles, books, websites and software tools about complex and social networks and about text mining in social media. |
|  |

|  |
| --- |
| There are fewer endeavours to characterize these networks beyond general features such as the scale-free |
|  |

|  |
| --- |
| aspect or to deal with text produced by social networks from the complex networks background. |
|  |

|  |
| --- |
| Research on network evolution is often restricted to network growth, in which there is a monotonic increase in the number of events~\cite{barabasiEvo}. |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Network types have been discussed with regard to the number of participants, intermittence of their activity and network longevity~\cite{barabasiEvo}. Two topologically different networks emerged from human interaction networks, depending on whether the frequency of interactions follows a generalized power law or an exponential connectivity distribution~\cite{barabasiTopologicalEv}. In email list networks, scale-free properties were reported with $\alpha \approx 1.8$~\cite{bird} (as in web browsing and library loans~\cite{barabasiHumanDyn}), and different linguistic traces were related to weak and strong ties~\cite{Gmane2}.  In this thesis, e-mail lists were chosen to represent human interaction networks, which were analyzed according to two broad aspects: the possible stability of the topology of ­the networks and the linguistic usage of distinct types of participants in the network. In the analysis, the topology of the networks was studied with regard to well-established topological metrics and well-established networks models. The text of the networks was studied with regard to statistics derived from the strings and from syntactics and semantics.   |  | | --- | | The fact that unreciprocated edges often exceed 50\% in human interaction networks~\cite{newmanEvolving} motivated the inclusion of symmetry metrics in our analysis. | |  |  |  | | --- | | No correlation of topological characteristics and geographical coordinates was found~\cite{barabasiGeo}, | |  |  |  | | --- | | therefore geographical positions were not considered in our study. | |  |  |  | | --- | | Gender related behavior in mobile phone datasets was indeed reported~\cite{barabasiSex} | |  |  |  | | --- | | but it is not relevant for the present work because email messages and addresses have no gender related metadata~\cite{gmanePack}. As for the language usage in the networks, emphasis was given to statistical features of the text such as number of tokens and known words, sizes of tokens, known words, sentences and messages.  Statistics were also derived from syntactics through POS tags (e.g. nouns, adjectives, verbs) and semantics through Wordnet synsets. |  |  | | --- | | The remainder of this chapter provides some background knowledge relevant to the topics discussed in the thesis, especially complex networks, text mining, graph visualization and linked data. In Chapter 2, the data and methods where described while the results and discussion are in Chapter~\ref{ch:disc}. | |  |  |  | | --- | | Conclusions and further work are stated in Chapter~\ref{ch:concl}. | |  |  |  | | --- | | The appendices bring further results from this research. | |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \section{Related knowledge} |
|  |

|  |
| --- |
| \subsection{Complex networks} |
|  |

|  |
| --- |
| Although not universally accepted, it is commonplace to define a complex network to be |
|  |

|  |
| --- |
| a ``graph with non-trivial topological features''~\cite{newmanBook}. |
|  |

|  |
| --- |
| We might add to this definition that a complex network is also a large graph (even though no consensus appears to exist as |
|  |

|  |
| --- |
| to what \emph{large} means in this context) |
|  |

|  |
| --- |
| and that it is a graph representation of a system found in natural, real or empirical system. |
|  |

|  |
| --- |
| Another way to approach the definition of ``complex networks'' is to define it as |
|  |

|  |
| --- |
| complex systems modeled as networks. |
|  |

|  |
| --- |
| This second definition is also useful but is even more problematic as |
|  |

|  |
| --- |
| there is no consensus of what a \emph{complex system} is. |
|  |

|  |
| --- |
| Even so, one should keep in mind that authors often define a complex system |
|  |

|  |
| --- |
| to be a system composed with many parts in which ``the whole is more than |
|  |

|  |
| --- |
| the sum of its parts''. |
|  |

|  |
| --- |
| Authors also often consider complex systems to have capabilities |
|  |

|  |
| --- |
| to ``process information'', to adapt and to reproduce~\cite{complexity}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| A graph is a structure that consists of a set of objects (called vertices) |
|  |

|  |
| --- |
| and a set of binary/dual relations of the objects (called edges). |
|  |

|  |
| --- |
| A graph might be unweighted and undirected (the simplest possibility), |
|  |

|  |
| --- |
| weighted and undirected, unweighted and directed, or weighted and directed. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The most usual representations of graphs (and networks) are the matrix, list and node-edge representations. |
|  |

|  |
| --- |
| In the matrix representation, each entry $a\_{ij}$ is non-zero if $i$ is linked to $j$; |
|  |

|  |
| --- |
| entries might be other than 0 and 1 in weighted graphs; undirected graphs yield symmetric matrices. |
|  |

|  |
| --- |
| There are two common list representations of graphs, one lists each pair of vertices that are connected, |
|  |

|  |
| --- |
| the other holds a list for each vertex in which are all the vertices connected to it (a list of lists). |
|  |

|  |
| --- |
| In the node-edge representation, each node is represented as a point while each edge is represented |
|  |

|  |
| --- |
| by a line between corresponding nodes. |
|  |

|  |
| --- |
| The matrix representation is essential for algebraic reasoning and for deriving metrics |
|  |

|  |
| --- |
| while the node-edge representation is important for illustration and intuitive guidance |
|  |

|  |
| --- |
| in characterizing the systems. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsubsection{A good justification for applying the complex networks theory} |
|  |

|  |
| --- |
| The estimated number of atoms in the universe, often used as a reference of largeness, |
|  |

|  |
| --- |
| is $\approx 10^{80}$. |
|  |

|  |
| --- |
| Let us find the number of vertices needed to reach such number of possible networks. |
|  |

|  |
| --- |
| Consider the simplest case of the unweighted and undirected networks. |
|  |

|  |
| --- |
| Each edge can exist or not (i.e. it is a Bernoulli variable) and with $n$ vertices there are |
|  |

|  |
| --- |
| at most ${n \choose 2}$ edges. |
|  |

|  |
| --- |
| Therefore: |
|  |

|  |
| --- |
| \begin{align} |
|  |

|  |
| --- |
| 2^{n \choose 2} > 10^{80} \Rightarrow |
|  |

|  |
| --- |
| log\_2[2^{n \choose 2}] > log\_2(10^{80}) \Rightarrow |
|  |

|  |
| --- |
| {n \choose 2} > \frac{log\_{10}(10^{80})}{log\_{10}2} \Rightarrow \nonumber\\ |
|  |

|  |
| --- |
| \Rightarrow \frac{n.(n-1)}{2} > \frac{80}{log\_{10}2} \Rightarrow |
|  |

|  |
| --- |
| N > 23,5988 \;\;\;\;\;\;\;\;\;\;\;\;\;\;\;\;\;\;\;\;\; |
|  |

|  |
| --- |
| \nonumber |
|  |

|  |
| --- |
| \end{align} |
|  |

|  |
| --- |
| That is, with only 24 vertices we have more possible networks than |
|  |

|  |
| --- |
| the estimated number of atoms in the universe. |
|  |

|  |
| --- |
| We should also add that the number of possible networks grows |
|  |

|  |
| --- |
| very fast with the number of vertices. |
|  |

|  |
| --- |
| This is a good reason for characterizing such systems by means |
|  |

|  |
| --- |
| of paradigmatic networks and generic metrics for nodes and the network (and edges, but it is less often). |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsubsection{Basic metrics}\label{se:intMea} |
|  |

|  |
| --- |
| Section~\ref{measures} gives a mathematical account of the following metrics, |
|  |

|  |
| --- |
| which are used here for characterizing basic types |
|  |

|  |
| --- |
| of networks in the next section. |
|  |

|  |
| --- |
| Such metrics are: |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item Degree $k\_i$: number of edges linked to vertex $i$. |
|  |

|  |
| --- |
| \item In-degree $k\_i^{in}$: number of edges ending at vertex $i$. |
|  |

|  |
| --- |
| \item Out-degree $k\_i^{out}$: number of edges departing from vertex $i$. |
|  |

|  |
| --- |
| \item Strength $s\_i$: sum of weights of all edges linked to vertex $i$. |
|  |

|  |
| --- |
| \item In-strength $s\_i^{in}$: sum of weights of all edges ending at vertex $i$. |
|  |

|  |
| --- |
| \item Out-strength $s\_i^{out}$: sum of weights of all edges departing from vertex $i$. |
|  |

|  |
| --- |
| \item Betweenness centrality $bt\_i$: fraction of geodesics that contain vertex $i$. |
|  |

|  |
| --- |
| \item Clustering coefficient $cc\_i$: fraction of pairs of neighbors of $i$ that are linked, i.e. the standard clustering coefficient metric for undirected graphs. |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
| % distance between vertices |
|  |

|  |
| --- |
| In the following discussion, we also use the concept of distance between a pair of nodes, |
|  |

|  |
| --- |
| which is the number of edges between the nodes. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsubsection{Basic types of networks} |
|  |

|  |
| --- |
| Complex networks are often characterized in terms of paradigmatic models. |
|  |

|  |
| --- |
| There are diverse models, but we can glimpse the background theory |
|  |

|  |
| --- |
| with the following ones~\cite{luMeasures}: |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item The Erdös-Rényi model\footnote{This name is also used for the model in which, for a fixed number of nodes and a fixed number of links, all networks are equally likely. This is the model originally introduced by Paul Erdös and Alfréd Rényi~\cite{erdosOrig}. |
|  |

|  |
| --- |
| We choose the definition given online (não seria inline, por ser uma definição dada no corpo principal do texto?), which is closely related to the one given in this footnote, because it is more commonly used nowadays.}: each pair of nodes is connected with a fixed random probability $p$. |
|  |

|  |
| --- |
| This model presents a characteristic degree ($n.p$ where $n$ is the number of nodes), low clustering and low average distance between nodes. |
|  |

|  |
| --- |
| \item Spatial network, also called geographic network or geometric graph: nodes are located in a metric space and the probability that two nodes are connected is greater as the distance between nodes gets smaller. These networks present characteristic degrees, high clustering and large average distance between nodes. |
|  |

|  |
| --- |
| \item Small-word network: defined as a network where the typical distance between nodes grows with the logarithm of the number of nodes while the average clustering coefficient is not small (larger than e.g. in the Erdös-Rényi model). |
|  |

|  |
| --- |
| One method for constructing a small-world network is to start with a regular lattice in which each node is connected to $k$ nearest neighbors. |
|  |

|  |
| --- |
| Each link is then rewired with probability $p$. |
|  |

|  |
| --- |
| With intermediate values of $p$ such as $0.01<p<0.1$, we obtain a network with both short average distance between nodes (as in the Erdös-Rényi model) and a high average clustering coefficient (as in the spatial network). |
|  |

|  |
| --- |
| This model presents a characteristic degree. |
|  |

|  |
| --- |
| \item Scale-free networks: in which the degree distribution $p(k)$ follows a power law ($p(k)=C.k^{-\alpha}$ where $C$ and $\alpha$ are constants). |
|  |

|  |
| --- |
| These networks are qualitatively characterized by the presence of a large number of poorly connected and of few highly connected hubs. |
|  |

|  |
| --- |
| Important is the absence of a characteristic degree, thus the name 'scale-free network'. |
|  |

|  |
| --- |
| \item Other networks: among important models of networks are exponential networks, networks with community structure and hybrid models. |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
| Real networks most often exhibit scale-free and small-world properties. |
|  |

|  |
| --- |
| This is the case of most e.g. social, gene and food networks. |
|  |

|  |
| --- |
| However, one should be cautious about such statement because |
|  |

|  |
| --- |
| the networks derived from the real systems depend heavily |
|  |

|  |
| --- |
| in what is considered a node and a link, |
|  |

|  |
| --- |
| i.e. on how the system is modeled as a graph. |
|  |

|  |
| --- |
| Another noteworthy remark is that |
|  |

|  |
| --- |
| the Erdös-Rényi networks, i.e. graphs of the Erdös-Rényi model, are frequently pin-pointed as the networks with trivial |
|  |

|  |
| --- |
| topological properties. |
|  |

|  |
| --- |
| This model is considered as a paradigmatic ``complex network'', concept often defined as graphs with non-trivial topological properties, |
|  |

|  |
| --- |
| which is a contradiction. Therefore, the term complex networks is not a very well defined notion, as it occurs for the \emph{complexity theory} in general. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Text mining of social data} |
|  |

|  |
| --- |
| Text mining is a multidisciplinary field, |
|  |

|  |
| --- |
| it is an extension of data mining to (often unstructured) textual data |
|  |

|  |
| --- |
| with the goal of discovering structure and meaning~\cite{customText}. |
|  |

|  |
| --- |
| A general outline of a text mining endeavor involves structuring input text, |
|  |

|  |
| --- |
| deriving patterns and the evaluation of the output. |
|  |

|  |
| --- |
| There are actually numerous models of such outline, |
|  |

|  |
| --- |
| as e.g. considering document collection and obtaining a final report in the |
|  |

|  |
| --- |
| start and end respectively~\cite{textSurvey}. |
|  |

|  |
| --- |
| Text mining tasks include document summarization, sentiment analysis |
|  |

|  |
| --- |
| and natural language processing techniques such as part of speech tagging~\cite{ntlk}. |
|  |

|  |
| --- |
| Among the applications one may include social media monitoring, automated ad placement, and development of tools for |
|  |

|  |
| --- |
| semantics ~\cite{textSurvey}. |
|  |

|  |
| --- |
| It is believed that applying text mining to social media |
|  |

|  |
| --- |
| can yield interesting findings in human behavior~\cite{customText}. |
|  |

|  |
| --- |
| Although there is no clear cut, text mining is sometimes divided into linguistic and non-linguistic~\cite{customText}. |
|  |

|  |
| --- |
| In the first case, techniques borrowed from linguistics are present, such as |
|  |

|  |
| --- |
| the analysis of discourse and part of speech tagging, |
|  |

|  |
| --- |
| and it is often mingled with natural language processing or computational linguistics (see Section~\ref{amb} for a coherent distinction of the fields). |
|  |

|  |
| --- |
| In the non-linguistic text mining, text is analyzed by means of statistical features |
|  |

|  |
| --- |
| derived from e.g. the size of tokens and sentences, |
|  |

|  |
| --- |
| and might be more easily related to the intuitive concept of data mining of text. |
|  |

|  |
| --- |
| In this thesis we use both perspectives. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Visualization of static and dynamic graphs} |
|  |

|  |
| --- |
| Static graph visualization is achieved in many ways, |
|  |

|  |
| --- |
| most usually through the node-link (often called network diagram) |
|  |

|  |
| --- |
| and matrix representations. |
|  |

|  |
| --- |
| Representing graphs as node-link diagrams has a long tradition which goes back |
|  |

|  |
| --- |
| at least to the works of Ramon Llull in the 13th century~\cite{llull}. |
|  |

|  |
| --- |
| To glimpse at the theory involved in visualizing networks~\cite{eades}, we mention three aspects: |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item criteria for the quality of layouts might include the number of crossing edges or the area of the drawing relative to the closest distance between two vertices. |
|  |

|  |
| --- |
| \item Layout methods are derived e.g. by placing vertices in a circular fashion, by using the eigen vectors from a worked out variant of the adjacency matrix as coordinates, or by force-based methods. For large graphs, including a number of social networks, the force-based networks are reported as useful. Therefore, we illustrate this method with the simplest model we could find in the specialized literature. Let $f\_a$ be the attraction force, $f\_r$ is the repulsion force, $d$ is the distance between the vertices and $k$ is a constant. The model introduced by Fruchterman and Reingold~\cite{fr} defines the forces as: |
|  |

|  |
| --- |
| \begin{align} |
|  |

|  |
| --- |
| f\_a = \frac{d^2}{k}\\ |
|  |

|  |
| --- |
| f\_d = -\frac{k^2}{d} |
|  |

|  |
| --- |
| \end{align} |
|  |

|  |
| --- |
| On a computer software, one usually starts with a random layout and performs a number of iterations updating the position of nodes |
|  |

|  |
| --- |
| using these forces to obtain the intended force-directed layout. |
|  |

|  |
| --- |
| \item Graph drawings are often developed for specific applications e.g. in biology (as for protein and gene interactions), in social networks, in tree diagrams. |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The core difference of dynamic graphs to static graphs is that vertices and edges |
|  |

|  |
| --- |
| can be added and removed over time. |
|  |

|  |
| --- |
| If we define the static graph G as $G:=(V,E)$ where V are the vertices as E are the edges in G, |
|  |

|  |
| --- |
| a dynamic graph might be defined as $\Gamma:=(G\_1,G\_2,...,G\_n)$ where $G\_i:=(V\_i,E\_i)$ |
|  |

|  |
| --- |
| are static graphs and indices refer to a sequence of time steps $(t\_1,t\_2,...,t\_n)$. |
|  |

|  |
| --- |
| In dynamic graph visualization most usually graphs are represented as animated diagrams |
|  |

|  |
| --- |
| or charts based on a timeline~\cite{dynGraph}. |
|  |

|  |
| --- |
| In this thesis we make use of node-link diagrams of both static and dynamic graphs. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Linked (open) data} |
|  |

|  |
| --- |
| The fields of social network analysis and complex networks |
|  |

|  |
| --- |
| are widely researched. |
|  |

|  |
| --- |
| However, there is a lack of open datasets for benchmarking results, |
|  |

|  |
| --- |
| especially associated with the complex networks field, |
|  |

|  |
| --- |
| yielding diverse results from poorly related sources. |
|  |

|  |
| --- |
| Recently, a myriad of results have been reported which are based on |
|  |

|  |
| --- |
| diverse datasets most often not accessible to researchers other than the publishing authors. |
|  |

|  |
| --- |
| In this thesis we present resources for having open databases to provide |
|  |

|  |
| --- |
| the scientific community with a friendly and common repertoire. |
|  |

|  |
| --- |
| We chose to use the linked data technology and follow W3C best practices for publishing data. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Linked data refers to data published in the web in such a way that it is |
|  |

|  |
| --- |
| machine readable and complies with a set of best practices. |
|  |

|  |
| --- |
| The web of data is constructed with documents on the web |
|  |

|  |
| --- |
| such as the web of HTML documents. |
|  |

|  |
| --- |
| In practice, the idea of linked data can me summarized |
|  |

|  |
| --- |
| by 1) the use of RDF to publish data on the web and 2) the use of RDF |
|  |

|  |
| --- |
| links to interlink data from different sources. |
|  |

|  |
| --- |
| The web is expected to be interconnected and to grow by the systematic application of four |
|  |

|  |
| --- |
| steps~\cite{lee1}: |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item Use URIs to identify things~\cite{uri}. |
|  |

|  |
| --- |
| \item Use HTTP URIs. |
|  |

|  |
| --- |
| \item Provide useful information when an URI is accessed via HTTP. |
|  |

|  |
| --- |
| \item Provide other URIs in the description of resources so human |
|  |

|  |
| --- |
| and machine agents can perform discovery. |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The Linked Open Data~\cite{lod} builds an ever growing cloud of data, |
|  |

|  |
| --- |
| the global data space, which is usually |
|  |

|  |
| --- |
| conceived as centered around the DBPedia, a linked data representation |
|  |

|  |
| --- |
| of data from Wikipedia~\cite{dbpedia0,dbpedia}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsubsection{RDF} |
|  |

|  |
| --- |
| The Resource Description Framework (RDF), a W3C |
|  |

|  |
| --- |
| recommendation, is a model for data |
|  |

|  |
| --- |
| interchange. |
|  |

|  |
| --- |
| It is based on the idea of making statements about resources in the form |
|  |

|  |
| --- |
| of triples, i.e. expressions in the form ``subject - predicate - |
|  |

|  |
| --- |
| object''. |
|  |

|  |
| --- |
| RDF can be serialized in several file formats, including RDF/XML, |
|  |

|  |
| --- |
| Turtle and Manchester, all of which, in essence, represent a labeled and |
|  |

|  |
| --- |
| directed multi-graph. |
|  |

|  |
| --- |
| RDF may be stored in a type of database referred to as a triplestore~\cite{rdf}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| As an example of an RDF statement, the following triple in the Turtle |
|  |

|  |
| --- |
| format asserts that ``the paper has color white'':\\ |
|  |

|  |
| --- |
| \texttt{http://example.org/Things\#Paper http://example.org/hasColor\\ |
|  |

|  |
| --- |
| http://example.org/Colors\#White .} |
|  |

|  |
| --- |
| % This work makes available an open database with diverse provenance and |
|  |

|  |
| --- |
| % to furnish the scientific community with a friendly and common repertoire. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Integration and uniformity of access is obtained through linked data |
|  |

|  |
| --- |
| representation, as explained in Section~\ref{queries}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Social participation} |
|  |

|  |
| --- |
| A significant share of our endeavor was oriented towards social participation, |
|  |

|  |
| --- |
| i.e. to facilitate civil engagement in a community, most significantly in State affairs. |
|  |

|  |
| --- |
| More concretely, we published data from a social participation federal portal~\cite{losd}, |
|  |

|  |
| --- |
| applied complex networks and text mining criteria for resources recommendation~\cite{pnud3,pnud4,opa} and |
|  |

|  |
| --- |
| proposed a ranking algorithm for voted proposals in another federal participation portal~\cite{dialogaAlg}. |
|  |

|  |
| --- |
| Such works were performed within a United Nations Development Program consulting contract, |
|  |

|  |
| --- |
| in partnership with the Brazilian Presidency of the Republic and published publicly mainly as technical reports~\cite{pnud3,pnud4,opa}. |
|  |

|  |
| --- |
| This aspect of our research was important for maturing topics and understanding |
|  |

|  |
| --- |
| the extent to which they are applied in pragmatic contexts. |
| It is left mostly as subsidiary documents in this thesis to allow for an easy public access. |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Other} |
|  |

|  |
| --- |
| Given the multidisciplinary condition of our work and of the implied topics, |
|  |

|  |
| --- |
| many other fields of knowledge could be further explored in this introduction or |
|  |

|  |
| --- |
| the methods chapter. |
|  |

|  |
| --- |
| To name just a few of the most directly related fields: statistics, principal component analysis, |
|  |

|  |
| --- |
| big data, social network analysis, social media mining, mathematical sociology, |
|  |

|  |
| --- |
| datasets, free culture, open source software, computer programming. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Of particular relevance are the typologies for human personality, such as the ones derived |
|  |

|  |
| --- |
| from the Myers-Briggs type indicator~\cite{mb} and from authoritarian personality~\cite{adorno} theories, |
|  |

|  |
| --- |
| because we present a new typology of human participants in social networks in Section~\ref{sec:pty}. |
|  |

|  |
| --- |
| Another topic we should highlight is what we called ``anthropological physics''~\cite{anPhy,antphy2}: |
|  |

|  |
| --- |
| the observation of natural/physical laws in human social systems. |
|  |

|  |
| --- |
| The term should not be confused with physical anthropology, which is a synonym for biological anthropology, |
|  |

|  |
| --- |
| a subfield of anthropology concerned with the evolution of humans~\cite{phyAn}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \section{Polysemy and synonyms}\label{amb} |
|  |

|  |
| --- |
| In the context of complex networks, the words \emph{network} and \emph{graph} |
|  |

|  |
| --- |
| are often used interchangeably, although the word graph might refer to the |
|  |

|  |
| --- |
| mathematical structure of vertices and edges and the word network might refer to the |
|  |

|  |
| --- |
| real system being represented as a graph or to the graph obtained by means of representing a real system. |
|  |

|  |
| --- |
| Furthermore, the word graph can be used to refer to a \emph{graph of a function} (mathematics) or to an abstract datatype (computer science). |
|  |

|  |
| --- |
| This parallelism between network and graphs also apply to network visualization and graph visualization. |
|  |

|  |
| --- |
| One might add here the term \emph{graph drawing}, another synonym for the visualization of graphs, although the term |
|  |

|  |
| --- |
| seems to be more traditional in relation to the achievement of node-edge network diagrams. |
|  |

|  |
| --- |
| Evolutionary graph visualization or evolutionary network visualization are examples of variants of \emph{dynamic graph visualization}. |
|  |

|  |
| --- |
| The nomenclature of vertices and edges varies widely among interested fields (mathematics, physics, biology, sociology, etc). |
|  |

|  |
| --- |
| A vertex might be called e.g. a node, a point, an agent, an actor, a participant. |
|  |

|  |
| --- |
| An edge might be called e.g. a link, a bond, a relation, a tie, a connection. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| % There is also some ambiguity related to metrics that can be take with respect to each vertex (or, less often, to each edge), |
|  |

|  |
| --- |
| % such as degree and clustering. |
|  |

|  |
| --- |
| % Such terms might be also used for the average of the measure for the whole network. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The terms \emph{text mining}, \emph{natural language processing} and \emph{computational linguistics} |
|  |

|  |
| --- |
| are often used for similar endeavors. |
|  |

|  |
| --- |
| A distinction might be made in that text mining refers to data mining of text, |
|  |

|  |
| --- |
| natural language processing is concerned with the interactions between the computer and the human natural languages, |
|  |

|  |
| --- |
| and computational linguistics aims for statistical or rule-based modeling of natural language from a computational perspective. |
|  |

|  |
| --- |
| Such fields are multidisciplinary and there is no sharp distinction between them. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Examined as fields of knowledge, the \emph{linked data} and the \emph{semantic web} |
|  |

|  |
| --- |
| terms are often used without distinction. |
|  |

|  |
| --- |
| Tim Berners-Lee coined both terms: |
|  |

|  |
| --- |
| the semantic web was conceived as a web of data that can be processed by machines~\cite{lee0}, |
|  |

|  |
| --- |
| the expression linked data appeared in a 2006 design note about the Semantic Web project~\cite{lee1} |
|  |

|  |
| --- |
| and refers to structured data that emphasizes interlinking and usefulness through semantic queries. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \emph{Social participation}, \emph{social envolvement} and \emph{social engagement} are synonyms that |
|  |

|  |
| --- |
| refer to the participation of an individual or group in a community or society. |
|  |

|  |
| --- |
| In Brazilian Portuguese, \emph{controle social} can refer to the antagonist concepts of social participation |
|  |

|  |
| --- |
| or of a social control (played by the State or companies in the civil society). |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{More specific terminology problems in the complex networks field} |
|  |

|  |
| --- |
| Given that this thesis involves multidisciplinary and new knowledge, |
|  |

|  |
| --- |
| it might be of no surprise that the nomenclature is not very well defined. |
|  |

|  |
| --- |
| Here we pin-point some more specific conflicts that arise in the literature of complex networks |
|  |

|  |
| --- |
| to both exemplify this issue and to avoid some problems in interpreting the |
|  |

|  |
| --- |
| methods and results in this thesis: |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item The \emph{hubs} are, by the usual definition, the more connected vertices. |
|  |

|  |
| --- |
| In the context of the HITS (Hyperlink-Induced Topic Search) algorithm, |
|  |

|  |
| --- |
| for attributing centrality to vertices, most traditionally to web pages, |
|  |

|  |
| --- |
| the hubs are the vertices with greater out-degree |
|  |

|  |
| --- |
| (greater in-degree yield \emph{authorities}). |
|  |

|  |
| --- |
| \item In some contexts, the center of network is the collection of vertices whose |
|  |

|  |
| --- |
| maximum distance to other vertices is the radius (i.e. the minimum maximum difference between vertices). |
|  |

|  |
| --- |
| In the same framework, the periphery of a network is the collection of vertices whose |
|  |

|  |
| --- |
| maximum distance to other vertices is the diameter (i.e. the maximum distance between vertices). |
|  |

|  |
| --- |
| By extension, the intermediary might be regarded as the set of vertices that are not in the center or the periphery. |
|  |

|  |
| --- |
| These definitions yield fractions of members that do not agree with the literature with respect to hubs, intermediary and periphery. |
|  |

|  |
| --- |
| We present a suitable method for deriving such classification, in the sense that it fits the literature prediction, in Section~\ref{sectioning}. |
|  |

|  |
| --- |
| \item Lace, loop, selfloop and autoloop are terms used to designate an edge from a vertex to itself. |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \section{Historical note} |
|  |

|  |
| --- |
| The knowledge fields involved in this thesis are very recent. |
|  |

|  |
| --- |
| To point just the main areas, |
|  |

|  |
| --- |
| complex networks has emerged in the final years of the 1990s decade~\cite{newmanBook}; |
|  |

|  |
| --- |
| text mining first workshops were held in 1999~\cite{textMining}; |
|  |

|  |
| --- |
| as an independent field, graph drawing arose in the 1990s~\cite{dynGraph}; |
|  |

|  |
| --- |
| the term linked data was coined in 2006~\cite{lee1}. |
|  |

|  |
| --- |
| % fields are recent, point starting works and/or dates for the fields. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \section{Further considerations} |
|  |

|  |
| --- |
| An initial proposal of this research was to enable |
|  |

|  |
| --- |
| the use of complex and social networks scientific knowledge by the participant of the networks. |
|  |

|  |
| --- |
| This motivated the open software, data and texts produced, conferences attended, and the endeavors with |
|  |

|  |
| --- |
| the United Nations, Brazilian Presidency and civil parties. |
|  |

|  |
| --- |
| As this was a practical goal, we found by hands-on processes that |
|  |

|  |
| --- |
| many fields are deeply related to the subject, which reflected in the number of fields tackled in this thesis |
|  |

|  |
| --- |
| and related documents~\cite{pnud3,pnud4,opa,anPhy,antphy2,dialogaAlg,losd,versinus,kolmSmir}. |
|  |

|  |
| --- |
| Furthermore, we understand that the open software, texts, videos and processes provided |
|  |

|  |
| --- |
| by our work contributes for the popularization of the knowledge and technologies implied |
|  |

|  |
| --- |
| by the empowerment of civil individuals and groups through the management of the |
|  |

|  |
| --- |
| networks in which they exist. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |

|  |
| --- |
| % canonical chapters; stability through circular statistics and topological statistics (PCA, sectioning) |
|  |

|  |
| --- |
| % text mining through measures of sizes of tokens, sentences, messages |
|  |

|  |
| --- |
| % through kolmogorov-smirnov, POS tagging, PCA |
|  |

|  |
| --- |
| % visualization of network evolution through animations (and other gadgets) |
|  |

|  |
| --- |
| % linked data representation of data and ontological organization |
|  |

|  |
| --- |
| % Appendices: supporting tables and diagrams, related results: |
|  |

% kolmogorov-smirnov, ubiquity of inequality through power laws, list of UNDP products, list of software