UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

RENATO FABBRI

Redes complexas para o participante

São Carlos

2015

RENATO FABBRI

Redes complexas para o participante

Monografia apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para o Exame de Qualificação como parte dos requisitos para obtenção do título de Doutor em Ciências.

Área de concentração: Física Aplicada
Opção: Física Computacional
Orientador: Prof. Dr. Osvaldo Novais de Oliveira Jr.

São Carlos

2015

RESUMO

FABBRI, C. *Redes complexas para o participante*. Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2015.

As redes complexas formam uma das áreas mais ativas da física redente. Há esforços consideráveis para apresentar estes avanços ao público geral, mas tudo indica que poucos ou nenhum são voltados para o benefício do indivíduo que constitui estes sistemas. Ou seja, com um núcleo de conhecimento da área, e receitas para aproveitamento, fornece meios para o participante interagir e entender as redes nas quais ele se encontra. Este trabalho objetiva realizar tal tarefa por meio das redes sociais do participante. Verificamos que tais redes exibem uma estabilidade temporal de medidas temporais e dos tamanhos relativos dos setores conectivos básicos (hubs, intermediários, periféricos). Observamos uma diferenciação da produção de texto de cada setor básico. Também formalizamos as conceitualizações destas redes como OWL onde foi possível, principalmente as instâncias de participação dosial previstas por lei e implementadas como software. Software e dados foram disponibilizados e usados. Protocolos escolhidos para facilitar a integração de estruturas de diferentes procedências, para reutilização dos dados em outros trabalhos e pesquisas, e para o benefício público. Consequências conceituais requerem considerações antropológocas e estão sendo documentadas. Próximos passos são: melhor documentação e desenvolvimento do aparato em software, ontologias e dados; considerações tipológicas das propriedades físicas observadas nas redes de interação humana, com atenção aos outliers, às relações entre topologia do agente e texto produzido, e à ponte com a bagagem mais tradicional das ciências humanas no assunto.

Palavras-chave: Redes complexas. Redes sociais. Complexidade. Física antropológica. Dados ligados. Web semântica. Participação social. Mineração de texto. Processamento de linguagem natural.

Sumário

1
Introdução

Estudos sobre redes de interação humana foram iniciados bem antes dos computadores modernos, datam do século XIX, enquanto a fundação da "Análise de Redes Sociais"/ARS (*Social Network Analysis*/SNA) é geralmente atribuida ao psiquiatra Jacob Moreno na metade do século vinte (**?**). Com a crescente disponibilidade de dados relacionados à interação humana, a pesquisa destas redes tem aumentado continuamente. Contribuições podem ser encontradas em uma variedade de áreas, de ciências sociais e humanidades (**?**) a ciências sociais (**?**) e física (**?**, **?**), dada a natureza multidisciplinar do assunto. Uma das abordagens da perspectiva de uma ciência exata é representar a rede de interação como uma rede complexa (**?**, **?**), com a qual algumas características foram reveladas. Por exemplo, a topologia das redes de interação humana exibem um traço livre de escala, o que aponta para a existência de um pequeno número de hubs super conectados e um grande número de vértices pouco conectados.

Há um hiato de conhecimento e tecnologi entre o legado de redes complexas e o usufruto do participante. Este hiato é reativo, e há evidência de que conseguirá se manter como um ecossistema de conhecimento, tecnologia e empreendimento da sociedade em todas as suas escalas. Deve facilitar, por exemplo: elaboração e preparação de documentos, aquisição rápida de conhecimento, realização de empreitadas coletivas. Em geral: processos de coleta e difusão de informação (e bens).

Este trabalho apresenta uma confirmação deste cenário e avanços. Algumas estratégias foram selecionadas para verificar a aplicabilidade de conceitos de redes complexas para o benefício do participane. Em especial, experimentos muito simples parecem capazes de modificar estruturas sociais. Neste contexto, verificamos estabilidades temporais nas redes de interação humana, e expomos que os setores primitivos das redes (hubs, intermediários e periféricos) produzem textos bastante diferentes entre si. Este conhecimento é útil para uma tipologia não estigmatizante de participantes em redes de interação. A audiovisualização e interconexão de dados com arte e engenhocas em software deram suporte contínuo à pesquisa científica.

Aplições foram complementadas com a Presidência da República e o PNUD/ONU.

A próxima seção apresenta considerações gerais sobre a literatura. A Seção 2 é dedicada aos dados analizados. A Seção **??** contém os métodos usados para atingir os resultados, que são explicitados na Seção **??**. O cronograma de atividades e uma comparação entre afazeres planejados, em andamento e finalizados estão na Seção **??**. A monografia termina com as conclusões na Seção **??**, seguida de agradecimentos e bibliografia.

1.1 Revisão de literatura

A área das redes complexas é relativamente nova (≈ 25 anos) e a literatura apresenta definições divergentes da área em si. Uma definição que tem recebido aceitação crescente é da rede complexa como "um grafo grande com características topológicas não triviais". Esta definição é enganosa ao menos em três pontos. Primeiro, há redes de interesse com características topológicas triviais, como as redes de Erdös-Rényi e a Geográfica (**?**), ou as redes simples usadas para exemplos. Segundo, a definição falha ao não emitir a mensagem fundamental de que uma rede complexas não é somente uma estrutura matemática, um grafo isolado. As redes complexas de interesse são redes reais ou modelos idealizados para as entender. Além disso, não só grafos grandes são de interesse, mas grafos pequenos são comumente usados como exemplos de propriedades e extensão das estruturas maiores. Uma definição, ainda longe de perfeita, mas preferida neste trabalho, é considerar a área das redes complexas como interessada em "redes usualmente grandes, consideradas no, ou para consideração do, meio em que residem". This definition resolves both issues.

Os livros em geral apresentam um comum e poderoso repertório para a caracterização de sistemas complexos através de grafos. Talvez mais notáveis sejam:

- O arsenal de medidas: grau, força, betweenness centrality, coeficiente de clusterização, etc.
- Os paradigmas básicos de redes: Erdös-Rényi, geográfica, de mundo pequeno e livre de escala.
- A abordagem transdisciplinar para considerar o meio no qual a rede está inserida, ou

que implica na rede.

A literatura sobre análise de redes sociais (ARS, ou *SNA* para *Social Network Analysis*), por exemplo, pode ser frequentemente compreendida como redes complexas em sistemas sociais humanos.

Uma consideração cuidadosa dos livros e artigos lidos para esta pesquisa estão na Seção **??**. As seções a seguir (1.1.1 e 1.1.2) explicitam peculiaridades do jargão da área e considerações sobre as áreas secundárias.

1.1.1 Ambiguidades e sinonimos no jargão

A área de redes complexas é recente e conflui com diversas correntes científicas, como a física, a biologia e a sociologia. Assim, possui termos ambiguos e sinònimos.

Exemplos de ambiguidade e delimitações adotadas:

- Os vértices mais conectados são, por definição, chamados hubs da rede. O vértice mais conectado é chamado hub da rede. No contexto do algoritmo HITS, o que é bem comum, estes significados mudam: os hubs são os que possuem mais arestas saindo (grau de saída); as autoridades recebem as arestas, ou são referenciados por vários hubs e outras entidades.

- Há uma definição de centro e periferia com relação ao raio e diâmetro da rede (**?**, **?**). Por extensão os intermediários podem ser considerados os que não são centro nem periferia. Esta setorialização centro, intermediários e periferia gera frações que diferem do previsto pela literatura para as frações de hubs, intermediários e periféricos. Um método apropriado para realizar esta setorialização da rede, com resultados estáveis e significativos, consta na Seção **??**.

- etc

1.1.2 Processamento de linguagem natural, dados ligados, participação social

Os termos processamento de linguagem natural (PLN) e mineração de texto (MT) podem em geral serem substituídos um pelo outro. O termo PLN é preferido pois os intuitos da pesquisa são muito mais próximos aos intuitos da área: compreender como a linguagem verbal está sendo usada para significar.

Os termos web semântica e dados ligados em geral também podem ser substituídos um pelo outro. O primeiro salienta a rede de referenciamento dos dados, o segundo os dados referenciando-se. Principalmente na esfera acadêmica, a área é, salvo segunda ordem, sinônimo de dados em RDF via XML ou Turtle, ontologias OWL e máquinas de inferência.

A participação social é a incorporação da própria sociedade nos processos de governança da sociedade. Quase toda a participação social atual é indireta e presencial, com a população fornecendo diretrizes e indicadores para o setor público. A transparência tem sido cada vez mais presente, e o norte de "participação direta" (participação direta da sociedade civil na tomada de decisões pelo Estado) cada vez mais presente.

2
Materiais

2.1 O banco Gmane de dados públicos sobre listas de email (benchmark)

Mensagens de lista de email foram obtidas do arquivo Gmane (**?**), que consiste em mais de 20 mil listas de email e mais de 130 milhões de mensagens (**?**). Estas listas cobrem uma variedade de assuntos, em especial relacionados à tecnologia. O arquivo pode ser descrito como um corpus com metadados de emails, que incluem hora e lugar de envio, nome e email do remetente. O uso do GMANE para pesquisa científica é incidente no estudo de listas isoladas e de inovações lexicais (**?**, **?**).

2.2 Facebook, Twitter, Participa.br, Cidade Democrática, AA

Embora as redes de email tenham sido usadas como referência na observação de propriedades gerais, outras fontes foram analisadas:

- Redes de amizade e interação do Facebook. 8 são usadas como referência em (**?**), mas dezenas, talvez algumas centenas, foram observadas nos experimentos da Seção 5.1.
- Milhares de tweets (talvez alguns milhões), geralmente vinculados à alguma *hashtag*. Em especial, a rede de retweets de 22 mil tweets com a hashtag #arenaNETmundial, foi analizada em (**?**).

- Mecanismos participativos como o Participa.br, Cidade Democrática e o AA. As redes de amizade e de interação do Participa.br foram analizadas em (**?**).

3
Métodos

Para realização desta pesquisa, foram necessários métodos consagrados, adequações e variantes. Esta seção expõe uma seleção destes métodos, para organizar o conhecimento e exemplificar esta diversidade:

- A Seção **??** expõe medidas simples de estatística circular, ou direcional. A contribuição neste caso é unicamente nos padrões encontrados, o método é bastante estabelecido.
- A Seção 3.1.1 expõe a síntese de redes de interação. Talvez haja contribuição na síntese do conceito de redes de interação, pois não encontramos (ainda) na literatura tal exposição concisa. De qualquer forma, o conceito e o procedimento para obtenção das redes a partir de dados é usual, a exposição neste texto e no artigo (**?**) serve principalmente ao intuito de formalização do processo.
- A Seção 3.2 é dedicada ao "Seccionamento de Erdös", para obtenção dos três setores básicos da rede, compostos por: hubs, intermediários e periféricos. O método parece não ter sido aplicado antes para este fim, e é resultado imediato da observação das caudas longas de dados reais contrastadas com a rede Erdös-Rényi (**?**).

3.1 Estatística temporal e circular

Para observação de padrões temporais, foram consideradas escalas diferentes. Em cada escala, de segundos e meses, foram construídos histogramas de atividade e feitas algumas medidas de estatística circular. A fração $\frac{b_h}{b_l}$ entre a maior b_h e a menor b_l incidência nos histogramas serviram como pista sobre quão uniforme são as distribuições observadas.

Considere cada $medida$ (dado pontual) como um número complexo com módulo 1, $z = e^{i\theta} = \cos(\theta) + i\sin(\theta)$, onde $\theta = medida\frac{2\pi}{periodo}$. Os momentos m_n, tamanhos dos momentos

R_n, ângulo médio θ_μ, e o ângulo médio reescalado θ'_μ são definidos assim:

$$\begin{aligned} m_n &= \frac{1}{N}\sum_{i=1}^{N} z_i^n \\ R_n &= |m_n| \\ \theta_\mu &= Arg(m_1) \\ \theta'_\mu &= \frac{period}{2\pi}\theta_\mu \end{aligned} \tag{3.1}$$

θ'_μ é usado como medida de localização. A dispersão é medida usando a variância circular $Var(z)$, o desvio padrão circular $S(z)$, e a dispersão circular $\delta(z)$:

$$\begin{aligned} Var(z) &= 1 - R_1 \\ S(z) &= \sqrt{-2\ln(R_1)} \\ \delta(z) &= \frac{1 - R_2}{2R_1^2} \end{aligned} \tag{3.2}$$

Como esperado, há uma correlação positiva entre $Var(z)$, $S(z)$ e $\delta(z)$, como pode ser notado nas informações de suporte de (**?**). A medida $\delta(z)$ foi preferida na discussão dos resultados.

3.1.1 Formação das redes de interação

Redes de interação podem ser modeladas tanto com quanto sem peso, tanto dirigida quando não dirigida (**?**, **?**, **?**, **?**). Neste trabalho, quando possível, consideramos redes dirigidas e com peso, a mais informativa das possibilidades. Nestes casos, desconsideramos as versões dirigidas sem peso, não dirigidas com peso e não dirigidas e sem peso.

Em geral, as redes de interação são obtidas da seguinte forma: uma reação direta do participante B a uma mensagem do participante A implica em uma aresta de A para B, representando a informação que foi de A para B. O raciocínio é: se B reagiu a uma mensagem de A, ele/ela leu o que A escreveu e formulou uma reação, portanto B assimilou informação de

A, assim $A \rightarrow B$. A inversão da direção da aresta produz a rede de status: B leu a mensagem e considerou o que A escreveu digno de resposta, dando status para A, portanto $B \rightarrow A$. Neste trabalho, as redes de interação são dirigida conforme o fluxo de informação, $A \rightarrow B$. A Figura 3.1 expõe esta formação. Maiores detalhes são:arestas em ambas as direções são consideradas distintas; selfloops são consideradas não informativas (para os interesses atuais) e descartadas; a primeira interação $A \rightarrow B$ cria a aresta com peso um; a cada nova interação $A \rightarrow B$ um é adicionado ao peso da aresta. Estas redes de interação humana constam na literatura como portadoras de propriedades livres de escala (e pequeno mundo), como esperado para (algumas) redes sociais (**?**, **?**).

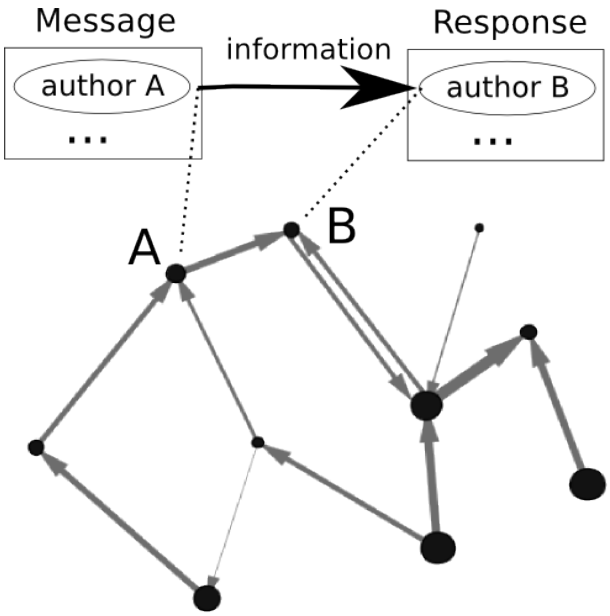

Figura 3.1 – A formação da rede de interação a partir de mensagens e respostas. Cada vértice representa um participante. Uma resposta do participante B a uma mensagem do participante A é considerada evidência de que B recebeu informação de A, representada então por uma aresta dirigida. Múltiplas mensagens adicionam "peso" à aresta dirigida. Maiores detalhes estão na Seção 3.1.1

3.2 Seccionamento de Erdös

Em uma rede livre de escala, os setores periféricos, intermediários e de hubs podem ser observados através de uma comparação com uma rede de Erdös Rényi com o mesmo número de arestas e vértices (**?**), como na Figura 3.2. Referiremos-nos a este procedimento como *seccionamento de Erdös*, com os setores resultantes chamados *setores de Erdös* (ou *setores*

primitivos, *setores básicos* da rede).

A distribuição de grau

The degree distribution $\widetilde{P}(k)$ of an ideal scale-free network $\mathcal{N}_f$ with N vertices and z edges has less average degree nodes than the distribution $P(k)$ of an Erdös-Rényi network with the same number of vertices and edges. Indeed, we define in this work the intermediary sector of a network to be the set of all the nodes whose degree is less abundant in the real network than on the Erdös-Rényi model:

$$\widetilde{P}(k) < P(k) \Rightarrow \text{k is intermediary degree} \tag{3.3}$$

If $\mathcal{N}_f$ is directed and has no self-loops, the probability of an edge between two arbitrary vertices is $p_e = \frac{z}{N(N-1)}$. A vertex in the ideal Erdös-Rényi digraph with the same number of vertices and edges, and thus the same probability p_e for the presence of an edge, will have degree k with probability:

$$P(k) = \binom{2(N-1)}{k} p_e^k (1-p_e)^{2(N-1)-k} \tag{3.4}$$

The lower degree fat tail represents the border vertices, i.e. the peripheral sector or periphery where $\widetilde{P}(k) > P(k)$ and k is lower than any intermediary sector value of k. The higher degree fat tail is the hub sector, i.e. $\widetilde{P}(k) > P(k)$ and k is higher than any intermediary sector value of k. The reasoning for this classification is: 1) vertices so connected that they are virtually inexistent in networks connected at pure chance (e.g. without preferential attachment) are correctly associated to the hubs sector. Vertices with very few connections, which are way more abundant than expected by pure chance, are assigned to the periphery. Vertices with degree values predicted as the most abundant if connections are created by pure chance, near the average, and less frequent in the real network, are classified as intermediary.

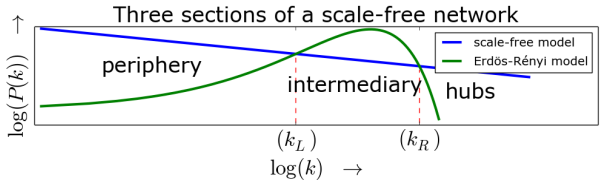

Figura 3.2 – Degree distribution of scale-free and Erdös-Rényi ideal networks. The latter has more intermediary vertices, while the former has more peripheral and hub vertices. The sector borders are defined by the two intersections k_L and k_R of the connectivity distributions. Characteristic degrees are in the compact intervals: $[0, k_L]$, $(k_L, k_R]$, $(k_R, k_{max}]$ for the Erdös sectors (periphery, intermediary and hubs).

i To ensure statistical validity of the histograms, bins can be chosen to contain at least η vertices of the real network. Thus, each bin, starting at degree k_i, spans $\Delta_i = [k_i, k_j]$ degree values, where j is the smallest integer with which there are at least η vertices with degree larger than or equal k_i, and less than or equal k_j. This changes equation 3.3 to:

$$\sum_{x=k_i}^{k_j} \widetilde{P}(x) < \sum_{x=k_i}^{k_j} P(x) \Rightarrow \text{i is intermediary} \tag{3.5}$$

If strength s is used for comparison, P remains the same, but $P(\kappa_i)$ with $\kappa_i = \frac{s_i}{\overline{w}}$ should be used for comparison, with $\overline{w} = 2\frac{z}{\sum_i s_i}$ the average weight of an edge and s_i the strength of vertex i. For in and out degrees (k^{in}, k^{out}) comparison of the real network should be made with:

$$\hat{P}(k^{way}) = \binom{N-1}{k^{way}} p_e^k (1-p_e)^{N-1-k^{way}} \tag{3.6}$$

where *way* can be *in* or *out*. In and out strengths (s^{in}, s^{out}) are divided by $\overline{w}$ and compared also using $\hat{P}$. Note that p_e remains the same, as each edge yields an incoming (or outgoing) edge, and there are at most $N(N-1)$ incoming (or outgoing) edges, thus $p_e = \frac{z}{N(N-1)}$ as with the total degree.

In other words, let γ and ϕ be integers in the intervals $1 \leq \gamma \leq 6$, $1 \leq \phi \leq 3$, and each of the basic six Erdös sectioning possibilities $\{E_\gamma\}$ have three Erdös sectors $E_\gamma = \{e_{\gamma,\phi}\}$ defined as:

$$\begin{aligned} e_{\gamma,1} &= \{\ i \mid \overline{k}_{\gamma,L} \geq \overline{k}_{\gamma,i}\} \\ e_{\gamma,2} &= \{\ i \mid \overline{k}_{\gamma,L} < \overline{k}_{\gamma,i} \leq \overline{k}_{\gamma,R}\} \\ e_{\gamma,3} &= \{\ i \mid\ \ \overline{k}_{\gamma,i} < \overline{k}_{\gamma,R}\} \end{aligned} \tag{3.7}$$

where $\{\overline{k}_{\gamma,i}\}$ is:

$$
\begin{aligned}
\overline{k}_{1,i} &= k_i \\
\overline{k}_{2,i} &= k_i^{in} \\
\overline{k}_{3,i} &= k_i^{out} \\
\overline{k}_{4,i} &= \frac{s_i}{\overline{w}} \\
\overline{k}_{5,i} &= \frac{s_i^{in}}{\overline{w}} \\
\overline{k}_{6,i} &= \frac{s_i^{out}}{\overline{w}}
\end{aligned}
\tag{3.8}
$$

and both $\overline{k}_{\gamma,L}$ and $\overline{k}_{\gamma,R}$ are found using $P(\overline{k})$ or $\hat{P}(\overline{k})$ as described above.

Since different metrics can be used to identify the three types of vertices, compound criteria can be defined. For example, a very stringent criterion can be used, according to which a vertex is only regarded as pertaining to a sector if it is so for all the metrics. After a careful consideration of possible combinations, these were reduced to six:

- Exclusivist criterion C_1: vertices are only classified if the class is the same according to all metrics. In this case, vertices classified (usually) do not reach 100%, which is indicated by a black line in Figure **??**.

- Inclusivist criterion C_2: a vertex has the class given by any of the metrics. Therefore, a vertex may belong to more than one class, and the total number of members may exceed 100%, which is indicated by a black line in Figure **??**.

- Exclusivist cascade C_3: vertices are only classified as hubs if they are hubs according to all metrics. Intermediary are the vertices classified either as intermediary or hubs with respect to all metrics. The remaining vertices are regarded as peripheral.

- Inclusivist cascade C_4: vertices are hubs if they are classified as so according to any of the metrics. The remaining vertices are classified as intermediary if they belong to this category for any of the metrics. Peripheral vertices will then be those which were not classified as hub or intermediary with any of the metrics.

- Exclusivist externals C_5: vertices are only hubs if they are classified as such according to all the metrics. The remaining vertices are classified as peripheral if they fall into

the periphery or hub classes by any metric. The rest of the nodes are classified as intermediary.

- Inclusivist externals C_6: hubs are vertices classified as hubs according to any metric. The remaining vertices will be peripheral if they are classified as such according to any metric. The rest of the vertices will be intermediary vertices.

Using equations 3.7, these compound criteria C_δ, with δ integer in the interval $1 < \delta < 6$ can be described as:

$$
\begin{aligned}
C_1 &= \{c_{1,\phi} = \{i \mid i \ \in e_{\gamma,\phi},\ \forall\, \gamma\}\} \\
C_2 &= \{c_{2,\phi} = \{i \mid \exists\ \gamma : i \in e_{\gamma,\phi}\}\} \\
C_3 &= \{c_{3,\phi} = \{i \mid i \ \in e_{\gamma,\phi'},\ \forall\, \gamma,\ \forall\, \phi' \geq \phi\}\} \\
C_4 &= \{c_{4,\phi} = \{i \mid i \ \in e_{\gamma,\phi'},\ \forall\, \gamma,\ \forall\, \phi' \leq \phi\}\} \\
C_5 &= \{c_{5,\phi} = \{i \mid i \ \in e_{\gamma,\phi'},\ \forall\, \gamma, \\
&\qquad\qquad \forall\ (\phi'+1)\%4 \leq (\phi+1)\%4\}\} \\
C_6 &= \{c_{6,\phi} = \{i \mid i \ \in e_{\gamma,\phi'},\ \forall\, \gamma, \\
&\qquad\qquad \forall\ (\phi'+1)\%4 \geq (\phi+1)\%4\}\}
\end{aligned}
\tag{3.9}
$$

The simplification of all the compound possibilities to the small set listed above can be formalized in strict mathematical terms, but this was considered out of the scope for current interests. It is worth noting that the exclusivist cascade is the same sectioning of an inclusivist cascade from periphery to hubs, but with inverted order of sectors precedence. These compound criteria can be used to examine network sections considering all degrees and strengths and are specially useful in the case of a low number of messages, such as in Section **??** of the Supporting Information.

3.3 Média e desvio do PCA ao longo do tempo

A Análise de Componentes Principais (PCA é a sigla consagrada, do inglês Principal Component Analysis) foi usada para observar a estabilidade na formação das componentes. Mais especificamente: 1) quais as medidas que contribuiem para cada componente e em que medida; 2) quanto da dispersão está concentrada em cada componente.

3.3.1 Medidas consideradas e acrescentadas

The topology of the networks was studied using Principal Component Analysis (PCA (**?**)) with a small selection of the most basic and fundamental measurements for each vertex, as follows:

- Degree k_i: number of edges linked to vertex i.
- In-degree k_i^{in}: number of edges ending at vertex i.
- Out-degree k_i^{out}: number of edges departing from vertex i.
- Strength s: sum of weights of all edges linked to vertex i.
- In-strength s_i^{in}: sum of weights of all edges ending at vertex i.
- Out-strength s_i^{out}: sum of weights of all edges departing from vertex i.
- Clustering coefficient cc_i: fraction of pairs of neighbors of i that are linked. The standard clustering coefficient for undirected graphs was used.
- Betweenness centrality bt_i: fraction of geodesics that contain vertex i. The betweenness centrality index considered directions and weight, as specified in (**?**).

In order to capture symmetries in the activity of participants, the following metrics were introduced for a vertex i:

- Asymmetry: $asy_i = \frac{k_i^{in} - k_i^{out}}{k_i}$.
- Mean of asymmetry of edges: $\mu_i^{asy} = \frac{\sum_{j \in J_i} e_{ji} - e_{ij}}{|J_i| = k_i}$, where e_{xy} is 1 if there is an edge from x to y, and 0 otherwise. J_i is the set of neighbors of vertex i, and $|J_i| = k_i$ is the number of neighbors of vertex i.
- Standard deviation of asymmetry of edges: $\sigma_i^{asy} = \sqrt{\frac{\sum_{j \in J_i} [\mu_{asy} - (e_{ji} - e_{ij})]^2}{k_i}}$.
- Disequilibrium: $dis_i = \frac{s_i^{in} - s_i^{out}}{s_i}$.
- Mean of disequilibrium of edges: $\mu_i^{dis} = \frac{\sum_{j \in J_i} \frac{w_{ji} - w_{ij}}{s_i}}{k_i}$, where w_{xy} is the weight of edge $x \to y$ and zero if there is no such edge.
- Standard deviation of disequilibrium of edges: $\sigma_i^{dis} = \sqrt{\frac{\sum_{j \in J_i} [\mu_{dis} - \frac{(w_{ji} - w_{ij})}{s_i}]^2}{k_i}}$.

3.4 Teste de Kolmogorov-Smirnoff para os textos produzidos por cada setor

Be $F_{1,n}$ and $F_{2,n'}$ two empirical distribution functions, where n and n' are the number of observations on each sample. The two-sample Kolmogorov-Smirnov test rejects the null hypothesis if:

$$D_{n,n'} > c(\alpha)\sqrt{\frac{n + n'}{nn'}} \tag{3.10}$$

where $D_{n,n'} = sup_x[F_{1,n} - F_{2,n'}]$ and $c(\alpha)$ is given for each level of α:

3.4.1 Adaptação

We need to compare empirical distribution functions, so $D_{n,n'}$ is given, as are n and n'. Therefore, as all terms in equation 3.10 are positive and $c(\alpha)$ can be isolated:

$$c(\alpha) < \frac{D_{n,n'}}{\sqrt{\frac{n+n'}{nn'}}} = c'(\alpha) \tag{3.11}$$

Tables **??**-**??** are populated with values for $c'(\alpha)$. When $c'(\alpha)$ is high, low values of α are possible for the test to reject the null hypothesis. Therefore, when $c'(\alpha)$ is greater than ≈ 1.7, it is reasonable to assume that $F_{1,n}$ and $F_{2,n'}$ differ.

3.5 Audiovisualização de dados

Networks were visualized with animations, image galleries and online gadgets developed specifically for this research (**?**,**?**,**?**). Such visualizations were crucial to guide research into the most important features of network evolution. Furthermore, the size of the three Erdös sectors could be visualized in a timeline fashion. Visualization of network structure was especially useful in the inspection of data and derived structures from the email lists.

3.6 Considerações tipológicas

Por enquanto, isso para nós quer dizer:

- Consideração sobre o potencial estigmatizante da classificação.
- Considerações sobre o indivíduo ou a rede que é classificada, i.e. recebe um tipo.

3.7 Web semântica

Para a formalização de conceitualizações, e para formatos de dados apropriados para armazenamento, compartilhamento e referência, foram adotadas as recomendações de linked data

da W3C (**?**, **?**, **?**).

3.7.1 A construção de ontologias OWL

Foram construídas ontologias e vocabulários a partir de entrevistas (OBS e VBS principalmente). Um método de levantamento de ontologia orientado aos dados surgiu, potencialmente útil a todos os portais e software em necessidade de ontologias, e foi responsável por 2 ontologias (OPA dados e OCD).

3.7.2 A triplificação de dados relacionais

Para disponibilização e uso de dados de diferentes fontes, foram feitos pequenos programas de computador (scripts) para acessar dados relacionais e escrever triplas RDF.

α	0.1	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

4
Resultados

4.1 Estabilidade temporal em redes de interação humana

Explicitados em (**?**).

4.2 Criação da nuvem brasileira de dados participativos

Primeiras ontologias, aplicações, dados, publicações, etc. Produtos 3,4,5 e artigo sobre OPS.

4.2.1 Conversão de dados relacionais para RDF

Roteiros de triplificação do Participa.br, Cidade Democrática, AA.

4.2.2 Síntese de ontologias OWL

Scripts para síntese, seja livre ou a partir dos dados.

4.2.3 Ontologias e dados publicados

datahub.io, MP, arXiv.

4.3 Aproveitamento

4.3.1 Procedimentos de percolação social

Procedimentos cíclicos e procedimentos efèmeros. Experimentos paradigmáticos:

- Crowdfunding para a pesquisa de redes (Vaca do fim do mundo).
- Escrita para os de maior betweenness e de maior closeness, mostrar que não bate para as listas.

4.3.2 Sistemas de recomendação para o enriquecimento da navegação semântica de recursos

Como descrito no produto 4 PNUD. Repassar caracterização básica e direcionar ao escrito.

4.3.3 Compreensão sobre a entidade social

Scale free as the consequence of T^2 signal, fractal, constant, with three primitive parts and greater specialization. Gradus ad Parnassum.

5
Afazeres e cronograma

5.1 Experimentos

5.2 Documents

5.2.1 To be finished

Anthropological physics

The study of human systems raises conceptual and ethical issues that require anthropological considerations. There are two immediate routes to this concepts:

- What data should or can be used?
- Can one experiment in a network of humans? In which context?

The short answer is that ethics committees and procedures are dedicated to dealing with those issues. Even so, there is a key-concept from the anthropological legacy: the study of the self as exposed to the interested culture or context. In this sense, it is reasonable (if not a suggestion) that a researcher do reflexive consideration, i.e. that he/she observe and make assumptions about its own sampling of the world. Within this same framework, many social networks (email, Facebook, Twitter, Participa.br, AA) were openly mined, with feedback to and from the studied communities. The term "anthropological physics" started being used in Brazil around 2014 and can be thought as a subfield of Social Physics.

Gradus

Uma lista detalhada de ambiguidades e sinônimos deverá completar o que está na Seção 1.1.1.

Fazer o $2^{(1002)}$

Consider a idealized constitution of these networks:

- the resources of the environment are the persons, each with an amount of time available.
- The amount of resource employed by the environment to the network is constant through all connective sectors

5.2.2 Finished

5.3 Cronograma

	2013		2014		2015	
Atividade	1°	2°	1°	2°	1°	2°
1	[•]	[•]	•	•		
2	[•]	[•]	[•]	[ ]	[ ]	[ ]
3	[ ]	[•]	[•]	[•]	[•]	[ ]
4	[ ]	[•]	[•]	[•]	[•]	[•]
5					[•]	[•]
6	[•]	[•]	[•]	[•]	[•]	[•]
7	[•]	[•]	[•]	[•]	[•]	[•]

Tabela 5.1 – Cronograma de atividades ao longo dos semestres, descritas na Seção 5.3. A marcação • indica previsão feita no início do doutorado. A marcação [ ] se refere ao relato e previsão, agora no final do 1° semestre de 2015. As principais diferenças do previsto foram: as disciplinas foram terminadas no primeiro ano; a revisão da literatura, os acréscimos aos modelos atuais com o foco no participante da rede, e a implementação computacional, estas três atividades estão sendo realizadas constantemente e devem durara até pouco antes da entrega e defesa da tese.

Este projeto foi inicialmente dividido segundo as etapas a seguir e usadas como referência na Tabela 5.1:

1. Créditos Obrigatórios: cumprimento dos créditos obrigatórios em disciplinas, exigidas pelo programa de Doutorado do IFSC/USP.

2. Revisão da literatura.

3. Acréscimos aos modelos atuais com o foco no participante da rede.

4. Implementação computacional.

5. Escrita da tese.

6. Escrita e publicação dos resultados em artigos.

7. Trocas com pessoas externas, estabelecimento de colaborações.

Considerações sobre estes itens:

1. Foram cursadas as disciplinas de Processamento de linguagem natural, Mineração de dados, Visualização de dados e Web semântica. Dediquei um ano inteiramente às disciplinas. Estranhamente, fechei todas com B. No mestrado, fazia mais de 30 créditos na graduação, 4 disciplinas na pós, pesquisa, e fechei todas com A.

2. A literatura de para o trabalho proposto é ampla e este aprofundamento tem sido constante,

3. Os acréscimos aos modelos atuais tem tido o foco no participante da rede.

4. Há implementação computacional de provas de conceito, bibliotecas, rotinas básicas e rotinas para replicar resultados do grupo de pesquisa.

5. A escrita da tese pode tomar vários rumos: pode consistir de um conjunto de artigos ou de uma monografia final. Acho mais provável que seja um conjunto de artigos centrados no descrido na Seção 5.2.1.

6. Conseguimos finalizar um artigo (**?**). Há ao menos mais um em condições de publicação (**?**) e outro mais indiretamente relacionado sobre música (**?**). Além destes, há mais estes artigos no arXiv (**?**, **?**, **?**, **?**), todos referentes ao trabalho do doutorado. Foram publicados em revista internacional os artigos AA e Images/Vilson, ambos sem a colaboração do orientador.

7. Parte substancial do trabalho consistiu em experimentos de coleta e difusão de informação, o que disparou reuniões, visitas e colaborações. Este processo foi iniciado logo antes do doutorado e pode ser apreciado, por exemplo, pelas visitas a São Carlos de parcerios de pesquisa, pela integração do pesquisador ao grupo de pesquisa Nexus, vinculado ao CNPq, e ao aporte do PNUD/ONU dado ao pesquisador, sobre o qual a Presidência da República se posicionou como beneficiária.

6
Conclusões

Há, a princípio, uma confirmação de que os conhecimentos de redes complexas possuem aplicações diversas e potencialmente benéficas para o participante. Por exemplo, os experimentos apresentaram modificações da estrutura social para comportar a pesquisa, e podem ser usados para comportar outros empreendimentos. Os estudos de estabilidade e diferenciação em redes de interação humana apontam na direção de tipologias de redes e de participantes, com base nos setores de Erdös. Um legado de dados ligados e abertos é conveniente para apresentar estes dados às comunidades acadêmicas e interessadas nas aplicações, para o qual foram adiantadas ontologias, vocabulários, rotinas de conversão de dados relacionais em RDF e os dados em si.

Uma direção simples é focar no *Complex Networks Gradus ad Parnassum*, que é um manual para introdução à àrea através do benefício das redes complexas para o indivíduo pela ação dele próprio. Uma via menos pedagógica, mas mais fundamental, é explorar as estabilidades encontradas: até que número de agentes a distribuição dos setores e a formação do PCA se mantém?; como caracterizar a intermitência dos agentes enquanto a distribuição grau é estável?

Há, entretanto, considerações teóricas profundas sobre a área, com implicações sobre a própria constituição das redes complexas. Ao mesmo tempo, os métodos utilizados também já parecem novos, e está faltando uma capacidade de apresentar isso para as revistas. Talvez por falta de equipe adequada para realizar o trabalho, talvez por falta de prática do doutorando na escrita e encaminhamento acadêmico.