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 (**?**, **?**).

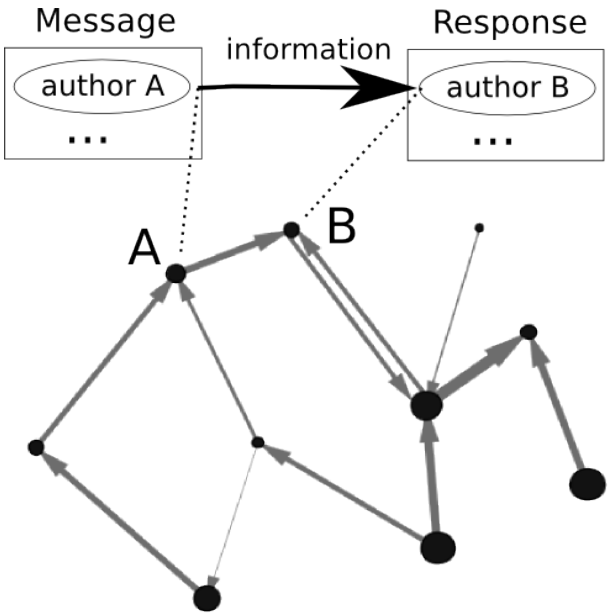

Figura 3.1 – A formação da rede de interação a partir de mensagens e respostas. Cada vértice representa um participante. Uma resposta do participante B a uma mensagem do participante A é considerada evidência de que B recebeu informação de A, representada então por uma aresta dirigida. Múltiplas mensagens 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 $\widetilde{P}(k)$ de uma rede livre de escala ideal $\mathcal{N}_f$ com N vértices e z arestas possui menos vértices com grau médios do que a distribuição $P(k)$ de uma rede Erdös-Rényi com o mesmo número de vértices e arestas. De fato, definimos (neste trabalho) o setor intermediário de uma rede como sendo o conjunto de todos os vértices cujo grau é menos abundante em uma rede real do que no modelo de Erdös-Rényi:

$$\widetilde{P}(k) < P(k) \Rightarrow \text{k é grau intermediário} \tag{3.3}$$

Se $\mathcal{N}_f$ for dirigida e não possuir selfloops, a probabilidade de existência de uma aresta entre dois vértices arbitrários é $p_e = \frac{z}{N(N-1)}$. Um vértice em um dígrafo de Erdös-Rényi com o mesmo número de vértices e aresta, portanto mesma probabilidade p_e para existência de aresta, terá grau k com probabilidade:

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

A cauda longa de graus baixos consiste nos vértices de borda, i.e. o setor periférico ou periferia, onde $\widetilde{P}(k) > P(k)$ e k é mais baixo que qualquer valor intermediário de k. A cauda londa de grau alto é o setor dos hubs, i.e. $\widetilde{P}(k) > P(k)$ e k é maior que qualquer valor de k do setor intermediário. O raciocínio para esta classificação é: os vértices tão conectados que são virtualmente inexistentes em redes conectadas por puro acaso (i.e. sem ligação preferencial) são corretamente associadas aos hubs. Vértices com pouquíssimas conexões, e muito mais abundantes do que esperado por puro acaso, são atribuídos à periferia. Vértices com valores de grau previstos como os mais abundantes caso as conexões sejam fruto de puro acaso, valores próximos da média, e menos abundantes em nas redes reais, são classificados como intermediários.

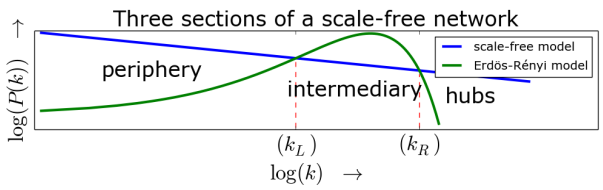

Figura 3.2 – As distribuições de grau de modelos ideais de redes livres de escala e Erdös-Rényi A segunda possui mais vérices intermediários, enquanto a primeira possui mais vértices periféricos e hubs. As bordas dos setores são definidas pelas duas intersecções k_L e k_R das distribuições de conectividade. Os graus característicos estão nos intervalos compactos: $[0, k_L]$, $(k_L, k_R]$, $(k_R, k_{max}]$ para os setores de Erdös (periferia, intermediários e hubs).

Para assegurar a validade estatística dos histogramas, os intervalos podem ser escolhidos de forma que contenham ao menos η vértices da rede real. Assim, cada intervalo, começando no grau k_i, estende-se por $\Delta_i = [k_i, k_j]$, onde j é o menor inteiro tal que há ao menos η vértices com grau maior que ou igual a k_i, e menos que k_j. Isso altera a equação 3.3 para:

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

Se a força s for usada para comparação, P permanece a mesma, mas $P(\kappa_i)$ com $\kappa_i = \frac{s_i}{\overline{w}}$ deve ser usado na comparação, com $\overline{w} = 2\frac{z}{\sum_i s_i}$ o peso médio da aresta e s_i o peso do vértice i. Para graus de entrada e saída (k^{in}, k^{out}) a comparação com a rede real deve ser feita com:

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

where *way* (sentido) pode ser *in* or *out* (entrada e saída). Forças de entrada e saída (s^{in}, s^{out}) são divididas por $\overline{w}$ e comparadas também usando $\hat{P}$. Note que p_e permanece a mesma, pois cada aresta é uma aresta de entrada (ou de saída), e há no máximo $N(N-1)$ arestas entrando (ou saindo), portanto $p_e = \frac{z}{N(N-1)}$ assim como no caso do grau total

Em outras palavras, seja γ e ϕ inteiros no intervalos. $1 \leq \gamma \leq 6$, $1 \leq \phi \leq 3$, e cada uma das seis possibilidades de seccionamento de Erdös $\{E_\gamma\}$ possui três setores de Erdös $E_\gamma = \{e_{\gamma,\phi}\}$ definidos como:

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

onde $\{\overline{k}_{\gamma,i}\}$ é:

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

e ambos $\overline{k}_{\gamma,L}$ e $\overline{k}_{\gamma,R}$ são encontrados usando $P(\overline{k})$ ou $\hat{P}(\overline{k})$ como descrito acima.

Como métricas diferentes podem ser usadas para identificas os três tipos de vértices, critérios compostos podem ser definidos. Após uma inspeção cuidadosa das possibilidades, os critérios compostos foram reduzidos a 6: $\{C_\delta\}_{\delta=1}^{6}$. Utilizando as Equações 3.7, estes critérios compostos C_δ, com δ inteiro no intervalo $1 \leq \delta < 6$ podem ser descritos como:

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

No artigo (**?**), os critérios C_1, C_3 e C_5 foram chamados exclusivistas, os critérios C_3 e C_4 de cascata e os critérios C_5 e C_6 de externos. Note que uma cascata exclusivista C_3 é a mesma classificação que uma cascata invertida (considera-se dos periféricos aos hubs) e inclusivista. Estes critérios compostos são especialmente úteis para observar estruturas com poucos participantes ou fruto de pouca atividade (veja as figuras do documento de Supporting Information de (**?**)).

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 principais. A PCA é bastante estabelecida e bem documentada e foi usado para saber: 1) quais as medidas que contribuiem para cada componente e em que proporção; 2) quanto da dispersão está concentrada em cada componente.

Ou seja, foram analizados os autovetores e autovalores das matrizes de vértices e suas medidas da seguinte forma: seja $\mathbf{X} = \{X[i,j]\}$ a matriz de todos os vértices i e respectivos valores de cada medida j, $\mu_X[j] = \frac{\sum_j X[j]}{J}$ a média da métrica j, $\sigma_X[j] = \sqrt{\frac{(X[j]-\mu_X[j])^2}{J}}$ o desvio padrão da métrica j, e $\mathbf{X}' = \frac{X[i,j]-\mu_X[j]}{\sigma_X[j]}$ a matriz com *z-score* de cada métrica j de de $\mathbf{X}$ em cada coluna. Seja $\mathbf{V} = \{V[j,k]\}$ a matriz JxJ de autovetores da matriz $\mathbf{C}$ de covariância de $\mathbf{X}'$, um autovetor por coluna. Cada autovetor combina as medidas originais em uma componente principal, portanto, basta observar $V'[j,k] = 100 * \frac{|V[j,k]|}{\sum_{j'} |V[j',k]|}$ para saber com que percentagem a medida j contribuiu para a componente principal k. Com o vetor de k autovalores $D[K]$, basta observar $D'[k] = 100 * \frac{D[k]}{\sum_{k'} D[k']}$ para saber a percentagem da dispersão pela qual a componente principal é responsável. Com os autovalores k ordenados de forma decrescente, em geral basta observar os primeiros três autovalores e respectivos autovetores em percentagens $\{(V'[j,k],\ D'[k])\}$, pois em geral já revelam padrões suficientes para uma boa análise e somam entre 60 e 95% da dispersão de todo o sistema. Em (**?**), em especial, foram feitas médias e desvios das constribuições de cada componente para a dispersão e das medidas em cada componente. Ou seja, dadas L observações l, cada uma com k pares de autovalores e autovetores, são observadas, para cada medida, a média $\mu_{V'}[j,k]$ e desvio $\sigma_{V'}[j,k]$ da medida j na componente principal k, e a média $\mu_{D'}[k]$ e desvio $\sigma_{D'}[k]$ da contribuição da componente k na dispersão do sistema:

$$\begin{aligned}
\mu_{V'}[j,k] &= \frac{\sum_l^L V'[j,k,l]}{L} \\
\sigma_{V'}[j,k] &= \sqrt{\frac{(\mu_{V'} - V'[j,k,l])^2}{L}} \\
\mu_{D'}[k] &= \frac{\sum_l^L D'[k,l]}{L} \\
\sigma_{D'}[k] &= \sqrt{\frac{(\mu_{D'} - D'[k,l])^2}{L}}
\end{aligned} \tag{3.10}$$

A matriz de covariância $\mathbf{C}$ também é observada diretamente para uma primeira pista sobre os padrões. Isso é feito com associações simples: valores absolutos pequenos indicam baixa correlação (a princípio independência); valores altos indicam correlação positiva (diretamente proporcional); valores negativos com módulo grande indicam correlação negativa (inversamente proporcional).

3.3.1 Medidas consideradas e acrescentadas

A topologia das rede sestudas foram estudadas utilizando PCA (**?**) com uma pequena seleção das medidas mais básicas e fundamentais de cada vértice. Formalmente, sejam i, j vértices e e_{ij} uma aresta de j para i (ou $j \to i$) e w_{ij} seu peso. Então:

- Grau $k_i = \sum_j (e_{i,j} + e_{j,i})$: número de arestas conectadas a i.
- Grau de entrada $k_i^{in} = \sum_j e_{i,j}$: número de arestas que terminam no vértice i.
- Grau de saída $k_i^{out} = \sum_j e_{j,i}$: número de arestas que partem do vértice i.
- Força $s = \sum_j (w_{i,j} + w_{j,i})$: soma dos pesos de todas as arestas conectadas ao vértice i.
- Força de entrada $s_i^{in} = \sum_j w_{i,j}$: soma dos pesos de todas as arestas que terminam no vértice i.
- Força de saída s_i^{out}: sima dos pesos de todas as arestas que partem do vértice i.

- Coeficiente de clusterização $cc_i = \frac{\sum e_{j_1 j_2}}{\binom{k_i}{2}}$: fração de pares de vizinhos j_1, j_2 de i que são conectados. A medida usual para grafos não direcionados foi usada.

- Intermediação (betweenness centrality) $bt_i = \frac{\Delta_i}{\Delta}$: fração entre o número Δ_i de geodésicas entre cada par de vértices da rede que contém o vértice i e Δ, o número total de geodésicas entre cada par de vértices da rede. A intermediação foi calculada considerando direções e peso, como especificado em (**?**).

Para apreender as simetrias das atividades dos participantes, as seguintes métricas foram introduzidas para o vértice i:

- Assimetria: $asy_i = \frac{k_i^{in} - k_i^{out}}{k_i}$.

- Média da assimetria das arestas: $\mu_i^{asy} = \frac{\sum_{j \in J_i} e_{ji} - e_{ij}}{|J_i| = k_i}$, onde e_{xy} é 1 se houver aresta de x para y, e 0 caso contrário. J_i é o conjunto de vizinhos do vértice i, e $|J_i| = k_i$ é o número de vizinhos do vértice i.

- Desvio padrão da assimetria das arestas: $\sigma_i^{asy} = \sqrt{\frac{\sum_{j \in J_i} [\mu_{asy} - (e_{ji} - e_{ij})]^2}{k_i}}$.

- Desequilíbrio: $dis_i = \frac{s_i^{in} - s_i^{out}}{s_i}$.

- Média do desequilíbrio das arestas: $\mu_i^{dis} = \frac{\sum_{j \in J_i} \frac{w_{ji} - w_{ij}}{s_i}}{k_i}$, onde w_{xy} é o peso da aresta $x \to y$ e zero se não houver tal aresta.

- Desvio padrão do desequilíbrio das arestas: $\sigma_i^{dis} = \sqrt{\frac{\sum_{j \in J_i} [\mu_{dis} - \frac{(w_{ji} - w_{ij})}{s_i}]^2}{k_i}}$.

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

Sejam $F_{1,n}$ e $F_{2,n'}$ duas distribuições cumulativas empíricas onde n e n' são o número de observações em cada amostragem. O teste de Kolmogorov-Smirnov de amostragem dupla rejeita a hipótese nula se:

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

onde $D_{n,n'} = sup_x[F_{1,n} - F_{2,n'}]$ e $c(\alpha)$ é dado para cada α segundo a Tabela 3.1:

Tabela 3.1 – Relação entre α e $c(\alpha)$ para o teste de Kolmogorov-Smirnov

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

3.4.1 Adaptação

Para alguns dos resultados, utilizamos $c(\alpha)$ como pista sobre o quão diferentes são pares de distribuições empíricas. São calculados $D_{n,n'}$, enquanto n e n' são dados. Assim, todos os termos da Equação 3.11 são positivos e $c(\alpha)$ pode ser isolado:

$$c(\alpha) < \frac{D_{n,n'}}{\sqrt{\frac{n+n'}{nn'}}} = c'(\alpha) \tag{3.12}$$

3.5 Audiovisualização de dados

Redes foram visualizadas com imagens, videos e engenhocas online para esta pesquisa (**?**, **?**, **?**). Também foram sonificadas, em especial como faceta auditiva de animações abstratas (**?**, **?**, **?**, **?**). Tais "audiovisualizações" foram cruciais para guiar a pesquisa para as características mais importantes das redes de evolução. Além disso, os tamanhos relativos dos três setores de Erdös foram visualizados como linhas temporais. A visualização da estrutura em rede foi especialmente útil na inspeção dos dados e estruturas das redes de email.

3.6 Considerações tipológicas e humanísticas

As redes estudadas são constituídos de seres humanos. Quando há classificação envolvida, sejam dos agentes ou dos sistemas em si, reflexões humanísticas são pertinentes, como as disparadas pelas perguntas:

- Qual o potencial estigmatizante da classificação?
- O que mais sabemos sobre o indivíduo ou a rede que é classificada, ou seja, é considerada(o) de um tipo?
- Quais dados posso usar sem desviar a atenção para leis e processos de comitês de ética?
- Qual a melhor forma de proceder com os dados e conhecimentos relacionados?

Todas estas questões, e muitas outras, estão em constante amadurecimento com grupos de pesquisa (**?**), leituras, e contatos individuais com outros pesquisadores (**?**, **?**).

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 dados ligados / web semântica da W3C (**?**, **?**, **?**). De forma bastante resumida, o arcabouço utilizado pode ser viso como uma forma de formalizar conceitos (classes), relações entre conceitos (propriedades) e instâncias dos conceitos (indivíduos). As informações são expressas de forma semi-estruturada em RDF: triplas "sujeito predicado objeto", com o sujeito sempre uma classe, o predicado sempre uma propriedade, e o objeto sempre uma classe ou dado. As propriedades podem ter aspectos específicos, chamados "axiomas de propriedade". As classes podem ser restritas a possuírem relações arbitrárias, chamadas "restrições de classe". É uma recomendação da W3C e o padrão acadêmico para dados ligados, i.e. para representação na web semântica.

Utilidades da tecnologia incluem:

- inferência por máquina através de especificações ontológicas.
- Interconexão de dados de fontes (bases de dados) diferentes.
- Organização de conhecimento específico para consideração cuidadosa, seja individual ou em grupo.

As ontologias são chave dentre as tecnologias de web semântica. Uma ontologia é geralmente definida como uma "especificação de uma conceitualização", e a recomendação é o uso do padrão OWL. Os vocabulários são coleções de termos e metadados, como definição, e a recomendação é o uso do padrão SKOS. O estado da arte de web semântica tem apresentado avanços, por exemplo, é capaz de realizar inferências úteis, especialmente para buscas. Por outro lado, é uma tecnologia complicada e com algumas dificuldades de implementação. Por exemplo, um conceito SKOS é um indivíduo, e uma classe OWL, se identificado com um conceito SKOS é, por consequência um indivíduo. Neste caso, os recursos de inferência por máquina ficam limitados dada a complexidade da especificação. Dito de outra forma, caso na especificação conste uma classe

3.7.1 A construção de ontologias OWL e vocabulários SKOS

Para formalizar conceitualizações referentes às estruturas sociais, mais especificamente relacionadas à participação social, foram construídas ontologias OWL e vocabulários SKOS a partir de entrevistas com especialistas: acadêmicos e gestores públicos. Também foram feitas ontologias e vocabulários a partir de bancos de dados, decretos presidenciais e outras documentações. O processo consistiu sempre que possível na coleta de informações, formalização dos conceitos e devolutiva aos entrevistados, com figuras e outras documentações, até que não tivessem mais contribuições.

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. Estes scripts criam conceitos novos e os vincula às instâncias dos dados. Na sequência, acessa as ontologias pertinentes, salva uma versão com os dados e ontologias, e uma versão com os dados, as ontologias e as triplas resultantes da inferência.

4
Resultados

4.1 Estabilidade temporal e diferenciação textual em redes de interação humana

Explicitados cuidadosamente em (**?**), os principais resultados são:

- A atividade ao longo do tempo é praticamente a mesma para todas as listas de email analisadas, e em todas as escalas. A maior dispersão foi encontrada nos segundos e minutos, seguida pelos dias do mês, meses, dias da semana e horas do dia. Padrões estáveis foram apreciados em todas estas escalas: segundos, minutos e dias do mês apresentaram uniformidade; meses parecem seguir calendários acadêmicos e escolares; dias da semana apresentam redução para dois ou um terço das atividades nos finais de semana; nas horas do dia, há concentração de atividades das 12-18h, o pico, porém, ocorre pouco antes das 12h.

- A fração de participantes em cada setor de Erdös é estável ao longo do tempo e pode ser determinada mesmo com poucas mensagens.

- As métricas topológicas se combinam nas componentes principais do PCA da mesma forma para todas as listas e todos os snapshots.

- As medidas de simetria da topologia, como definidas na Seção 3.3.1, apresentam mais dispersão do que o usual coeficiente de clusterização. O coeficiente de clusterização se combina com os desvios padrões de assimetria e desequilíbrio para a formação da terceira componente.

- Estes comportamentos são muito estáveis para em redes de interação de email. Nas outras redes analisadas, Twitter e Participa.br apresentaram redes bastante similares às

de email. Nas redes do Facebook foram encontradas algumas redes que diferiam do modelo apresentado pelas redes de listas públicas de email.

- Para um mesmo número de mensagens (sejam 20 mil) e diferentes listas, há uma correlação negativa entre número de participantes e número de threads quando os participantes são poucos (até 2 mil participantes quando são 20 mil mensagens). Para uma quantidade maior de participantes, há uma correlação positiva entre o número de participantes e o número de threads. Este fato deve estar relacionada a outras características topológicas e textuais da rede e pode servir para uma tipologia das próprias redes.

- Especulações humanísticas, especialmente sobre questões tipológicas e antropológicas, são seguem imediatamente os resultados quantitativos. Em especial, a setorialização de Erdös implica em uma tipologia de agentes em redes humandas de interação. Esta tipologia é, a princípio, não estigmatizante pois os agentes mudam de setor o tempo todo. Além disso, um mesmo agente pertence a todos os setores ao mesmo tempo, mas em redes diferentes. Maiores qualificações desta tipologia, decorrente do pertencimento a um setor de Erdös, estão no proprio artigo.

Com base nestes resultados, foi investigada a produção de texto na rede, com foco na potencial relação entre topologia, setor de Erdös e texto produzido (**?**). As principais conclusões são:

- Os textos produzidos por cada setor de Erdös diferem bastante: os $c(\alpha)$ fruto das comparações são tão grandes que as tabelas não registram os valores. Além disso, as diferenças entre setores iguais de redes diferentes são, via de regra, maiores que as encontradas entre setores diferentes da mesma rede.

- As características topológicas e textuais de cada agente apresentam correlações não triviais (como entre intermediação e uso de advérbios) e triviais (como entre grau e número de caractéres escritos). Mesmo assim, são muito menos correlacionadas entre si do que separadamente. Ou seja, as componentes principais possuem tendência a prevalência de medidas topológicas **ou** textuais, mas combinam-se medidas de ambos os tipos.

- Dada uma quantidade grande de texto, algumas medidas estatísticas relacionadas ao tamanho das palavras parecem se conservar. Não foi dada uma explicação definitiva

para este fenômeno, embora haja esperança de que consigamo explicações simples (o problema não parece difícil). Algumas destas medidas estão ao final de (**?**), resultantes de somas cumulativas de diferencas de histogramas.

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. Além disso, foi feita uma oficina, na Secretaria-Geral da Presidência da República, para explicitar a utilidade destas formalizações semânticas e coletar informações sobre diversos mecanismos e instâncias de participação social previstos em lei e praticados.

(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).

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.