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 compõem uma das áreas mais ativas da física recente. Há esforços consideráveis para apresentar estes avanços ao público não especialista, mas tudo indica que poucos ou nenhum são propostos para instrumentalizar o indivíduo que constitui estes sistemas a se beneficiarem. Ou seja, com um núcleo de conhecimento da área, e receitas para aproveitamento, fornecer meios para o participante interagir e entender as redes nas quais ele se encontra. Este trabalho busca realizar tal tarefa por meio das redes sociais do participante. Verificamos que tais redes exibem uma estabilidade temporal de medidas topológicas e dos tamanhos relativos dos setores básicos (hubs, intermediários, periféricos). Observamos uma diferenciação da produção de texto de cada setor básico. Também formalizamos as conceitualizações vinculadas a estas redes como ontologias OWL onde foi possível, principalmente as instâncias de participação social previstas por lei e praticadas ou implementadas como computacionalmente. 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 redigidas. 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 (ou *Social Network Analysis*/SNA) é geralmente atribuída 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 tecnologia 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 participante. 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 básicos 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. Aplicações foram complementadas em parceria com a Presidência da República e o Programa

das Nações Unidas para o Desenvolvimento.

A próxima seção apresenta considerações gerais sobre a literatura. A Seção 1.2 expõe a proliferação de ambiguidades e sinônimos no jargão da área. A Seção 2 é dedicada aos dados analisados. 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 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 complexa 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". Esta definição resolve ambas as questões.

Os livros em geral apresentam um comum e poderoso repertório para a caracterização de sistemas complexos através de grafos. Talvez as mais notáveis características deste repertório 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 (1.2 e 1.2.1) explicitam peculiaridades do jargão da área e considerações sobre as áreas secundárias.

1.2 Ambiguidades e sinônimos no jargão

A área de redes complexas é recente e conflui com diversas correntes científicas, como a física, a biologia e a sociologia. Portanto, possui termos ambíguos e sinônimos.

Exemplos de ambiguidade, sinônimos 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 **??**.

- *Aresta* e *ligação* são usadas como sinônimos. *Nó* e *vértice* também. É comum o uso de outros termos, em geral coerentes com a aplicação, como *agente*, *ator* ou *participante*

para vértices de redes observadas em sistemas humanos.

- Laços, *loops*, *selfloops*, *autoloop*, *buckle* são termos usados para arestas de um vértice para ele próprio.

1.2.1 Processamento de linguagem natural, dados ligados, participação social

Diversos títulos foram lidos sobre processamento de linguagem natural, mineração de texto, visualização de dados e web semântica. Estas áreas tem impacto sobre o que está feito, e sendo feito, e foram cursadas formalmente uma disciplina sobre cada uma para o doutorado. Seguem informações pontuais sobre cada área.

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 nesta pesquisa pois o intuito é mais confluente: 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 visualização de dados de grafos em evolução temporal é bastante incipiente. Os poucos casos da literatura foram visitados. As animações abstratas de redes em evolução, e as "audio-visualizações" das redes, que disponibilizamos como parte desta pesquisa, são potencialmente contribuições na fronteira da visualização. Vídeo, porém, não é o formato mais apreciado pela literatura de visualização de dados, que tende a qualificar as figuras bidimensionais como as mais apropriadas para a pesquisa científica.

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, indicadores e acompanhamento para o setor público. A participação social tem sido fortalecida no mundo todo, e conceitos como transparência, participação direta

(participação direta da sociedade civil na tomada de decisões pelo Estado) e democracia líquida (atribuição recursiva de competência para tomada de decisão), se estabelecendo aos poucos como diretrizes para governos, acadêmicos e sociedade civil.

2
Materiais

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

Mensagens de listas de email foram obtidas através do arquivo Gmane (**?**), que consiste em mais de 20 mil listas de email e mais de 130 milhões de mensagens públicas (**?**). 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 4.4.2.
- 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 analisada em (**?**).

- Mecanismos participativos como o Participa.br, Cidade Democrática e o AA. As redes de amizade e de interação do Participa.br foram analisadas 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 3.1 expõe medidas simples de estatística circular, ou direcional. A contribuição neste caso é unicamente nos padrões encontrados, o método é bastante estabelecido.

- A Seção 3.2 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 3.3 é dedicada à "Setorialização 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 o modelo de Erdös-Rényi (**?**).

- A Seção 3.4 apresenta o uso mais recorrente da Análise de Componentes Principais (PCA) neste trabalho. Várias redes são observadas, ou a mesma rede é observada em vários momentos, e a concentração de dispersão das componentes principais, e das medidas nas componentes principais, são observadas através de médias e desvios padrão.

- A Seção 3.4.1 apresenta as medidas utilizadas nas análises, com exposição formal das medidas de simetria potencialmente novas (não encontramos ainda na literatura), mas bastante úteis nas análises.

- A Seção 3.5 apresenta o uso que fazemos do teste de Kolmogorov-Smirnov de amostragem dupla. O método é bem estabelecido, e a contribuição está nos resultados alcançados com ele sobre diferenciação da produção de texto nas redes de interação.

- A Seção 3.6 expõe sobre a utilização dos dados de redes sociais para geração de imagem, música, e animação abstrata.
- A Seção 3.7 explicita a pertinente recorrência de considerações qualitativas e do cânone das ciências humanas.
- A Seção 3.8 delineia um pouco as abordagens utilizadas para registrar conceitualizações e vinculá-las aos dados.

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. Considere cada $medida$ (dado pontual) como um número complexo de 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 e pode ser notado nas informações de suporte de (**?**), há uma correlação positiva entre $Var(z)$, $S(z)$ e $\delta(z)$. A medida $\delta(z)$ foi preferida na discussão dos resultados. A fração $\frac{b_h}{b_l}$ entre a maior b_h e a menor b_l incidência nos histogramas também serviram como pista sobre quão próximas à distribuição uniforme são as distribuições observadas.

3.2 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; laços são consideradas não informativos (para os interesses atuais) e descartados; 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 (**?**, **?**).

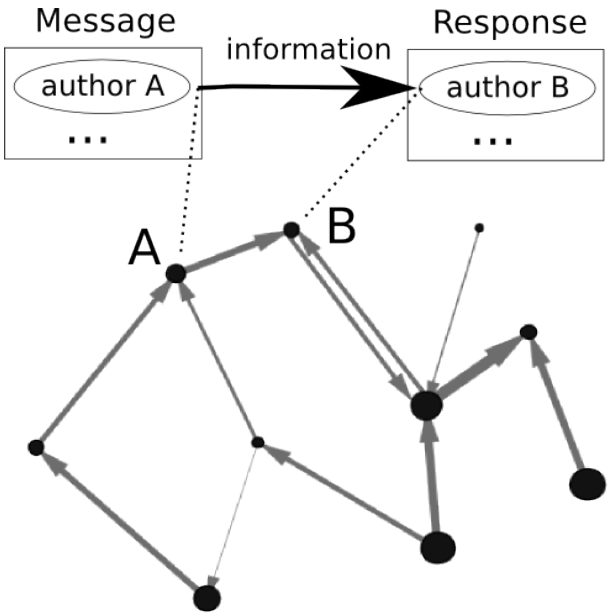

Figura 3.1 – A formação da rede de interação a partir de mensagens e respostas. Cada vértice representa um participante. Uma resposta do participante B a uma mensagem do participante A é considerada evidência de que B recebeu informação de A, representada então por uma aresta dirigida. Múltiplas mensagens adicionam "peso" à aresta dirigida. Maiores detalhes estão na Seção 3.2

3.3 Setorialização 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 *setorialização 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 médio 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 laço (aresta de um vértice para ele próprio), 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 ditada pela distribuição binomial:

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

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 longa 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.

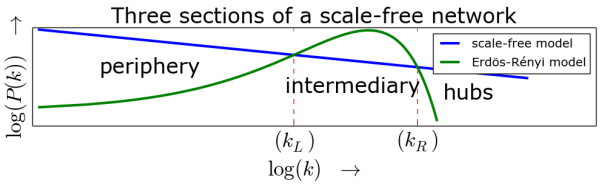

Figura 3.2 – As distribuições de grau de modelos ideais de redes livres de escala e Erdös-Rényi. A segunda possui mais vértices intermediários, enquanto a primeira possui mais vértices periféricos e hubs. As bordas dos setores são definidas pelas duas intersecções k_L e k_R das distribuições de conectividade. 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:

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

onde *way* (sentido) pode ser *in* ou *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, sejam γ e ϕ inteiros nos intervalos $1 \leq \gamma \leq 6$, $1 \leq \phi \leq 3$. Cada uma das seis possibilidades de setorialização 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 a medida γ no vértice i, convencionada:

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

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

Como métricas diferentes podem ser usadas para identificar 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 (**?**)).

3.4 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 usada para saber: 1) quais as medidas que contribuem para cada componente e em que proporção; 2) quanto da dispersão está concentrada em cada componente.

Ou seja, foram analisados 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 $J \times J$ 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 que percentagem da a componente principal k é contribuição da medida j. 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 contribuições 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).

3.4.1 Medidas consideradas e acrescentadas

A topologia das rede foram estudadas utilizando PCA (**?**) com uma pequena seleção das medidas mais básicas e fundamentais de cada vértice.

As seguintes medidas bastante conhecidas foram usadas: grau, grau de entrada, grau de saída, força, força de entrada, força de saída, coeficiente de clusterização, centralidade de intermediação (*betweenness centrality*). Além disso, 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: $\sigma_i^{dis} = \sqrt{\frac{\sum_{j \in J_i} [\mu_{dis} - \frac{(w_{ji} - w_{ij})}{s_i}]^2}{k_i}}$.

3.5 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' contam as observações em cada amostragem. O teste de Kolmogorov-Smirnov de amostragem dupla rejeita a hipótese nula (rejeita que $F_{1,n}$ seja fruto da mesma distribuição que $F_{2,n'}$) 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 região crítica α (probabilidade da hipótese nula ser verdadeira) segundo a Tabela 3.1.

Tabela 3.1 – Relação entre a região crítica α e $c(\alpha)$ para o teste de Kolmogorov-Smirnov de amostragem dupla.

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

3.5.1 Adaptação

São calculados $D_{n,n'}$, enquanto n e n' são dados. Todos os termos da Equação 3.11 são positivos e $c(\alpha)$ pode ser isolado:

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

Utilizamos $c'(\alpha)$ como distância entre pares de distribuições empíricas, o que é coerente com a teoria (**?**).

3.6 Audiovisualização de dados

Redes foram visualizadas com imagens, videos e engenhocas online para esta pesquisa (**?**, **?**, **?**). Também foram sonificadas, em especial como faceta sonora de animações abstratas (**?**, **?**, **?**, **?**). Tais "audiovisualizações" foram cruciais para guiar a pesquisa para características relevantes das redes de interação. 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.

3.7 Considerações tipológicas e humanísticas

As redes estudadas neste trabalho são constituídas por seres humanos. Quando há classificação envolvida, seja 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 e que procedimentos posso realizar sem desviar a atenção da pesquisa para leis e processos de comitês de ética?
- Qual a melhor forma de proceder com os dados e conhecimentos frutos da pesquisa?

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

3.8 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 maneira 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 especificidades , chamadas "axiomas de propriedade". As classes podem ser restritas a possuírem certas relações, 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 diferentes.
- Organização ontológica 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, as inferências tem se tornado mais ágeis e 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 (instância de uma classe), e uma classe OWL, se identificado com um conceito SKOS é, por consequência um indivíduo. Neste caso, quando um um indivíduo (instância de uma classe) é também uma classe, dada a complexidade, os recursos de inferência por máquina ficam limitados e lentificados.

3.8.1 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 (**?**).

3.8.2 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 com os dados semanticamente enriquecidos. Estes *scripts* formalizam conceitos e os vinculam aos dados. Na sequência, acessam as ontologias pertinentes, salvam uma versão com os dados e ontologias, e uma versão com os dados, as ontologias e as triplas resultantes da inferência sobre os dados com a ontologia.

4
Resultados

4.1 Estabilidade temporal e topológica; diferenciação textual em redes de interação humana

Explicitados cuidadosamente em (**?**), os principais resultados da estabilidade temporal e topológica em redes de interação humana 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 pela dispersão encontrada nos 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, mas o pico, porém, ocorre pouco antes das 12h.

- A fração de participantes em cada setor de Erdös é estável ao longo do tempo e esta estrutura já desponta na rede mesmo com poucas mensagens.

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

- As medidas de simetria da topologia, como definidas na Seção 3.4.1, são responsáveis por mais dispersão do que o coeficiente de clusterização. Resultado menor: 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 em dois aspectos: algumas proporções e combinações de medidas das componentes principais; frações de participantes em cada setor de Erdös.

- 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 $\approx$ 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, seguem imediatamente dos procedimentos e resultados quantitativos deste trabalho. Em especial, a setorialização de Erdös implica em uma tipologia de agentes em redes humanas de interação. Esta tipologia é, a princípio, não estigmatizante pois os agentes mudam de setor constantemente. 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 final dos resultados do 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:

- O texto produzido por cada setor de Erdös é bastante diferente um do outro: os $c(\alpha)$ fruto do teste de Kolmogorov-Smirnov entre histogramas de uso de recursos textuais (pontuação, adjetivos, etc) de cada setor são tão grandes que as tabelas não registram os valores (veja Seção 3.5). Além disso, as diferenças entre $c(\alpha)$ de setores iguais de redes diferentes são, na grande maioria das vezes, maiores que as encontradas entre setores diferentes de uma mesma rede. Isso decorre de uma maior discrepância de massa probabilística entre os histogramas de setores diferentes de uma mesma rede do que entre setores iguais de redes diferentes.

- 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 caracteres 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 a combinação de medidas de ambos os tipos é incidente.

Estes resultados permearam várias outras frentes de de pesquisa e desenvolvimento tecnológico (**?**, **?**, **?**, **?**, **?**, **?**, **?**).

4.2 Criação da nuvem brasileira de dados participativos ligados

Iniciada para formalizar as rede e participantes, junto às suas propriedades e estruturas. Esta frente rapidamente se voltou para as formalizações de conceitualizações referentes às estruturas e sistemáticas já em prática e previstas em lei. Dados também foram publicados, fazendo uso das ontologias feitas. Estes dados e ontologias foram em grande parte já publicados e estão em uso, mas a grande maioria não recebeu artigo científico ainda.

4.2.1 Síntese de ontologias OWL e vocabulários SKOS

Ontologias OWL feitas neste trabalho:

- OPS (Ontologia de Participação Social): é uma ontologia fruto de diversos esforços da América latina, principalmente do Brasil. Nesta pesquisa, revisamos a ontologia e disponibilizamos um código OWL em uso por instâncias diferentes da academia, Estado e sociedade civil.
- OPa (Ontologia do Participa.br): esta é uma ontologia feita para e partir dos dados do Participa.br (Portal Federal de Participação Social, SG-PR).
- OPP (Ontologia de Portais Participativos): esta ontologia foi pensada com a equipe do

Participa.br e outros especialistas como esquema geral de portais participativos. Ontologia relativamente complexa, centrada em 3 classes: Participante, Comunidade, Mecanismo Participativo.

- Ontologiaa (Ontologia do AA): é uma pequena ontologia para o minimalista AA (Autorregulação Algorítmica), um software para registrar e compartilhar processos intelectuais como para pesquisa e arte (**?**, **?**).

- OCD (Ontologia do Cidade Democrática): é uma ontologia extensa para o portal participativo Cidade Democrática, da sociedade civil. Dado o tamanho da ontologia, o processo de sua construção deu origem ao método de construção de ontologia OWL a partir dos dados, descrito na Seção **??** e utilizado também para a construção da OPa (acima).

- OBS (Ontologia da Biblioteca Social): Esta ontologia consiste na verdade em uma coleção de ontologias, uma para cada conceito que precisasse de uma abordagem dedicada, e uma para cada mecanismo ou instância de participação social prevista no Decreto Presidencial nº 8.243 de 23 de maio de 2014, conhecido como decreto da PNPS ou da Política Nacional de Participação Social. Esta ontologia também contou com entrevistas feitas diretamente para este trabalho, além da contribuição de uma oficina na 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.

O VBS (Vocabulário da Biblioteca Social) é uma adaptação (com complementos) da OBS no formato de vocabulário SKOS, principalmente para permitir uso junto ao DSPACE.

As ontologias e vocabulários são todas construídas através de scripts, com exceção da OPP, feita no Protegé.

4.2.2 Formatação de dados ligados a partir de dados relacionais

Três roteiros de conversão de dados foram feitos para conversão de dados relacionais em dados RDF enriquecidos semanticamente:

- Triplificação do Participa.br: foi elaborado e disponibilizado um roteiro para formatação em RDF de dados relacionais do Participa.br (originalmente em PostgreSQL) (**?**). Estes dados foram depois usados, através de buscas SparQL, para auxiliar na construção da OPa (**?**).

- Triplificação dos dados do Cidade Democrática: foi elaborado um roteiro para formatação em RDF de dados do Cidade Democrática (originalmente em MySQL) (**?**). Estes dados foram depois utilizados para auxiliar na construção da OCD (**?**).

- Triplificação dos dados do AA: foi elaborado um roteiro para formatação em RDF de dados do AA encontrados em bancos de dados MySQL e MongoDB, e em *logs* de IRC (**?**). Esta triplificação é especialmente útil pela simplicidade do AA, o que permite experimentar abordagens diferentes e escolher a melhor, antes de desenvolver ontologias e triplificações mais complicadas. Esta foi a única triplificação feita depois da ontologia e não aproveitada para a construção da ontologia (**?**).

4.2.3 Escritos, ontologias e dados publicados

Foi publicado no arXiv um artigo sobre a OPS (**?**). Este escrito aguarda confluência com orientador para publicação em revista, potencialmente na revista PLOS ONE. Foram escritos também os produtos PNUD/ONU, em processo de publicação em instâncias governamentais, mas já em repositórios públicos (**?**, **?**, **?**). Foram publicados dados em RDF do Participa.br, Cidade Democrática e AA no Datahub.io (**?**). Ontologias e vocabulários foram publicadas junto ao ministério do planejamento e em repositórios públicos (**?**, **?**, **?**, **?**, **?**) *Scripts* para síntese das ontologias e triplificação de dados estão também publicamente acessíveis e junto aos produtos PNUD citados acima.

4.2.4 Método de construção de ontologias orientado aos dados

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). Resumidamente, o método consiste em: representar os dados de interesse como RDF; realizar buscas SparQL para construir ontologia trivial com as classes e propriedades encontradas; realizar buscas SparQL para inferir restrições de classe e axiomas de propriedade (**?**, **?**, **?**).

4.3 Aparato em software

Scripts para verificar as estabilidades topológicas e diferenciações textuais em redes humanas estão reunidos em um pacote oficial da linguagem Python (**?**). Estão sendo feitos pacotes para organizar os numerosos *scripts* de triplificação de dados, construção de ontologias e vocabulários e mineração das estruturas (**?**, **?**). Os dados, classes e propriedades das ontologias e triplificações estão também disponíveis (em parte) através das próprias URIs, redirecionadas via purl.org para um servidor de pesquisa. Ou seja, caso você acesse `http://purl.org/socialparticipation/opa/Participant`, o servidor em http://purl.org redireciona seu navegador para um servidor de pesquisa com várias entidades do conceito "Participante" da ontologia "opa". Os dados estão em um *endpoint* SparQL, e *scripts* para a mineração destes dados estão disponíveis em interfaces web via um IPython Notebook. As ontologias estão também disponíveis na instalação do Webprotegé da Stanford (**?**). Muitas engenhocas foram criadas para gerar figuras, vídeos e inspecionar estruturas sociais de emails, Facebook, Twitter, Participa.br, AA, IRC e outras fontes (**?**, **?**, **?**, **?**). Outras engenhocas foram criadas para experimentações estéticas e informacionais (**?**, **?**, **?**, **?**).

4.4 Aproveitamento, utilidade e formalismo

A proposta do trabalho é comprovar a demanda e desenvolver algumas formas para o participante se beneficiar de suas redes através do legado das redes complexas. Os resultados das seções anteriores servem a este fim, sendo, porém, mais desenvolvimentos científicos e tecnológicos do que formas para o participante se beneficiar. Esta seção complementa a monografia neste aspecto: registra andamentos fronteiriços do trabalho, difíceis de formalizar e até inconclusivos, mas cuja utilidade para o participante é latente.

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

O relacionamento semântico de dados e conceitualizações via tecnologias de web semântica torna os recursos navegáveis à semelhança do que fazemos com os navegadores Web ao abrir páginas HTML (por isso a área chama-se **web** semântica). Ao invés de páginas HTML, os recursos são formatados em RDF e os links são consequência de critérios semânticos. No decorrer desta pesquisa, surgiram possibilidades de enriquecimento da navegação semântica através de recomendações de recursos com métodos abertos e propostas de aproveitamento pelo usuário. Esta versão recebeu até prova de conceito (**?**):

- São geradas estruturas auxiliares: rede de amizades, rede de interação, histograma de radicais (morfemas do texto), seleção dos 400 radicais mais incidentes para caracterizar o domínio, histograma de radicais de cada recurso (postagem, comentário, participante, etc.).
- O solicitante pode requerer recomendação de recursos de qualquer tipo a partir de um recurso de qualquer tipo. Pode optar pelo método de recomendação topológico (utilizando as redes de amizade e interação), textual (utilizando os histogramas de radicais) ou híbrido (utilizando ambos). Pode optar por polaridade de similaridade (recomenda recursos similares), dissimilaridade (recomenda recursos dissimilares) ou mista (mistura

de recursos similares e dissimilares ou recomendação na qual essa classificação não se aplica).

- Os métodos são todos explicitados em texto e código para o participante. Cada método conta também com um registro de potenciais utilidades para o participante, assim como cada recomendação.

4.4.2 Experimentos de percolação social

Foram realizados procedimentos cíclicos e procedimentos efêmeros de difusão de informação para observar as reações e testar hipóteses de modificação das estruturas sociais. Experimentos paradigmáticos:

- Em dezembro de 2012 foram iniciados ciclos de coleta e difusão de informação sobre as redes sociais e o potencial benéfico para o indivíduo civil. Duraram meses e redes diferentes foram usadas. Foram confirmadas as hipóteses de modificação das estruturas sociais para comportar a pesquisa, com suporte humano, financeiro e institucional. Foram confirmadas hipóteses de modificação do tratamento da sociedade sobre o tema. Foram confirmadas hipóteses de que seriam verificáveis estes resultados tanto em minhas interações cotidianas, via praticamente todos os meios. Estas foram algumas consequências da "percolação do tecido social" (mudança abrupta das propriedades físicas do tecido social acompanhado de mudança gradual de conectividade). Em especial, a minha rede de amizades do Facebook foi utilizada (cada vértice é um amigo meu, cada aresta indica uma amizade entre eles), e amigo por amigo foi acionado, dos menos conectados aos mais conectados.

- Analisando redes de amizade do Facebook, percebi que portavam uma característica não intuitiva: em praticamente qualquer rede de tamanho médio ou grande (mais de 500 pessoas), dentre as 50 pessoas com a maior intermediação (*betweenness centrality*, mais participa de geodésicas) haviam sempre pouquíssimas que constavam entre as 50 com maior closenness centrality (mais perto de todos os outros agentes). Com base nesta diferenciação inesperada, selecionei estes dois grupos em minha rede e em

redes de parceiros que também fazem experimentos semelhantes. Cada um de nós elaboramos uma mensagem diferente, e a mandamos para nossos respectivos dois grupos, separadamente. Ou seja, cada pessoa enviou uma mensagem diferente, cada grupo de cada pessoa recebeu uma cópia desta mensagem. O grupo com a maior intermediação reagia sempre calorosamente, repassava a mensagem, até interagiam entre si, mesmo sem se conhecerem ou serem próximos. Os grupos de maior *closenness* nunca reagiam, saíam rapidamente da interface usada para a mensagem. A hipótese mais plausível que surgiu para explicar esta diferença de reação é a de que os de maior intermediação tinham maior influência sobre a rede, enquanto os de maior *closenness* sofriam maior influência.

- Em um evento grande em São Paulo sobre transparência e governança na internet, foi operacionalizado um telão de streaming de estruturas sociais, escrito no percurso desta pesquisa, que expunha em tempo real as três diferentes redes de Twitter formada por usuários relacionados por *retweet*, vocabulário e *#hashtag*. A tecnologia pode ser compreendida como "*streaming* de estruturas sociais", e gerou bastante discussão com as pessoas que foram ao evento, inclusive com os próprios comunicadores que constavam nas redes de *retweet*. Houve confirmação da hipótese de que as pessoas se interessariam e se instruiriam. Houve alguns usos a mais da ferramenta, localizados e a pedido do meio, não por necessidade da pesquisa.

O segundo experimento foi feito por vários parceiros de pesquisa, por onde pôde ser verificado o comportamento constante. O primeiro experimento ainda não foi replicado. É comum após alguma apresentação ou reunião de pesquisa alguém se prontificar a fazê-lo, mas isso nunca aconteceu. Eu mesmo já me comprometi comigo a replicar o experimento, mas não aconteceu. Uma hipótese usual é que haja bloqueios mentais que nos impedem de realizar uma intervenção tão direta na nossa existência social, ou nosso eu-rede (**?**, **?**).

4.4.3 Física antropológica

Estes experimentos, e outras anotações de dados, são, no escopo deste trabalho, considerados questionáveis, potencialmente inapropriados, caso não sejam observadas algumas

diretrizes:

- estudo das redes das quais o pesquisador faz parte, como um estudo de si.
- Uso de anotações (de si) com a devida atenção para não expor as pessoas desnecessariamente e para quaisquer maiores cuidados sugeridos pelo contexto.
- Abertura constante dos procedimentos, dados, códigos e literatura produzida.

Estas diretrizes foram apreendidas em grande parte da tradição antropológica, e, portanto, configuram uma pesquisa com alguns aspectos"antropológicos". O termo "física antropológica" começou a ser usado no Brasil principalmente por acadêmicos (físicos, cientistas da computação, filósofos, antropólogos) em 2013-14, no contexto dos experimentos de difusão de informação e das análises, ambos em minhas próprias redes. Considerações cuidadosas estão sendo feitas constantemente sobre o presente trabalho, sobre o termo, sobre o legado antropológico, sobre a física e as redes complexas, e sobre termos relacionados, como física social (**?**) e sociofísica (**?**). Há resistência do meio científico, mas no geral o balanço apontada para uma pertinência do uso do termo para representar o que está sendo feito.

4.4.4 Compreensão sobre as estruturas sociais

Praticamente todos os resultados citados acima são úteis para compreender as estruturas sociais e são fruto desta tentativa. Há a intenção de disponibilizar um compêndio às redes complexas através da instrumentalização do leitor com estes conhecimentos e tecnologias para exploração de si próprio. Um esboço consta em (**?**).

Um exemplo especial de fundamentação que parece não constar na literatura é a constatação de que a propriedade livre de escala é consequência da distribuição equânime de recursos pelos setores da rede. Para apreender este fato, considere uma quantidade fixa R de recursos que será utilizada para a realização da rede em cada setor. Considere que, para cada quantidade de recursos T, são contadas $f = \frac{R}{T}$ partes de tamanho T, como na Figura 4.1.

Segue que $log(f) = -1log(T) + C$, $C = log(R)$ uma constante arbitrária. Uma reta descreve a relação entre $log(f)$ e $log(T)$, como na Figura **??**. Os recursos são alocados pelo sistema de forma uniforme, pois $T\frac{R}{T} = R = constante$.

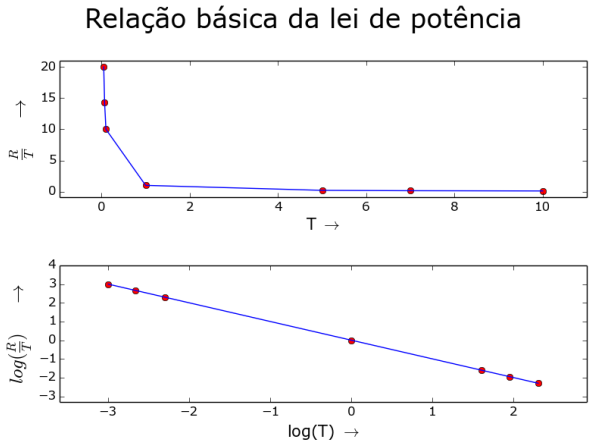

Figura 4.1 – A curva resultante da divisão de uma mesma quantidade R de recursos em $\frac{R}{T}$ partes de tamanho T. Utilizada para expor uma potencial causa da ubiquidade da propriedade livre de escala.

Considere que $T = T_1T_2$ (e.g. recursos da rede=agentes x tempo de cada agente). Neste caso, $f = \frac{R}{T_1T_2}$ e segue que $log(f) = -log(T_1T_2) + C$. Se $T_1 = T_2$, $log(f) = -2log(T_1) + C$, e $\gamma = 2$ como previsto pela literatura. No exemplo, o tempo alocado é o tempo dos próprios agentes, portanto é razoável considerar $T_1 = T_2$. Possíveis causas para a distorção do valor exato $\gamma = 2$ são: propriedades fractais, recursos em número diferente, associações entre os recursos.

5
Afazeres e cronograma

5.1 Cronograma

Tabela 5.1 – Cronograma de atividades ao longo dos semestres, descritas na Seção 5.1. 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 terminaram 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 durar até a entrega e defesa da tese.

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

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 Introdução ao Processamento de Língua Natural (SCC5908, 12 créditos), Mineração de Dados não Estruturados (SCC5920, 12 créditos), Visualização Computacional (SCC5836, 12 créditos), e Introdução à Web Semântica (SCC5929, 8 créditos). Por um ano as disciplinas foram priorizadas. No mestrado, fazia mais de 20 créditos na por semestre graduação, 6 disciplinas na pós em um ano (66 créditos) e pesquisa, e fechei todas com A. Estranhamente, no doutorado fechei todas as disciplinas com B, fiz menos disciplinas na pós, não fiz graduação e desprendi tempo para as disciplinas.

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, em especial através das tipologias e considerações antropológicas.

4. Há implementações computacionais de provas de conceito, bibliotecas, rotinas básicas e rotinas para replicar resultados do grupo de pesquisa. Engenhocas para gerar arte audiovisual a partir de redes.

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 focados nas direções dadas na Seção 4.4.4.

6. Conseguimos finalizar um artigo (**?**). Há ao menos mais dois em condições de publicação (**?**, **?**). 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 parceiros 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 (veja a Seção 4.4.2 e (**?**) para mais detalhes).

5.2 Comparativo de afazeres

Tabela 5.2 – Relação de tarefas feitas e por fazer. Há literatura pronta e vários documentos escritos e em mãos para serem aprofundados. A literatura a ser consumida é bastante extensa e cursos em vídeo, como do Coursera, são vistos com frequência (alguns cursos foram vistos inteiros). O mais urgente parece fazer uma revisão e aprofundamento de estatística e física estatística. Falta confirmar os experimentos percolatórios contínuos (veja Seções 4.4.2 e 4.4.3). As disciplinas foram terminadas.

	feito	**por fazer**
escrita	artigo de estabilidade em redes de interação humana (**?**); artigo sobre a Ontologia de Participação Social (**?**); ensaio descrevendo simbiose com PNUD/ONU e SG-PR (**?**); artigo com descrição psicofísica da música no áudio digital (**?**); produtos PNUD 3, 4 e 5, descrevendo sistemas de classificação, recomendação, ontologias e triplificações para participação social com métodos de redes complexas e processamento de linguagem natural (**?**, **?**, **?**); artigo sobre AA (**?**); versões iniciais e rascunhos dos artigos sobre física antropológica (**?**), sobre votação contínua por aprovação e participação (**?**), sobre diferenças da produção textual nos setores de Erdös (**?**), sobre visualização de redes de interação em evolução temporal (**?**), sobre audiovisualização de redes de interação em evolução temporal (**?**), sobre performance audiovisual via controle coletivo de código e projeção ao vivo (**?**)	publicar artigos no arXiv; repassar produtos PNUD um e dois; "Complex Networks Gradus ad Parnassum", um compêndio de redes complexas que utiliza a existência em rede do leitor para instrumentalizá-lo; artigo sobre tipologia de agentes humanos em redes de interação; versão desenvolvida do escrito sobre física antropológica; documentação do pacote Python oficial "percolation" (**?**); artigo com o método de levantamento de ontologias orientado aos dados; artigo sobre os dados participativos ligados brasileiros; artigo com os experimentos de coleta e difusão de informação; versão final do ensaio do AA (**?**)
leitura	documentação de redes complexas; documentação de web semântica; amadurecimentos coletivos frutos das difusões de informação; numerosos artigos da Wikipédia, protocolos e manuais de software; cursos do Coursera, alguns completos; literatura de PLN; literatura de visualização de dados e mineração de dados; artigos, exemplos especiais são ()	estatística e física estatística, talvez manuais de R também; terminar livros referência de redes complexas; absorver uma literatura mínima sobre antropologia; visita à topologia tradicional e teoria de grafos na computação
experimentos	experimentos contínuos/cíclicos e outros efêmeros	confirmar experimentos contínuos/cíclicos
comunidade	repassados resultados para comunidades estudadas; confirmada permissão dos desenvolvedores do Gmane para utilizar os dados das listas para pesquisa	repassar às comunidades estudadas um resumo dos resultados, em linguagem mais acessível que os artigos
disciplinas	cursadas disciplinas Introdução ao Processamento de Linguagem Natural, Mineração de dados; Visualização de dados; Introdução à web semântica	-//-
considerar banca	-//-	preparar apresentação; apresentar e anotar contribuição da banca; conduzir com orientador

Tabela 5.3 – Relação de tarefas feitas e por fazer. Há bastante software pronto, principalmente para experimentos e provas de conceito, mas também para pesquisa científica. A finalização de pacotes oficiais da linguagem Python está planejada. Foram publicados também dados, ontologias e vocabulários relacionados à participação social. Arte audiovisual é feita junto às visualizações de dados.

	feito	**por fazer**
software	telões de streaming de estruturas sociais; funcionalidades escolhidas da MMISSA (Monitoramento Massivo e Interativo da Sociedade pela Sociedade para Aproveitamento); engenhoras no AARS (A Análise de Redes Sociais) e MyNSA (*Monitoring yields Natural Streaming and Analysis*); rotinas de triplificação de dados; rotinas de construção de ontologias; rotinas para, dada a rede social, sintetizar música e animação visual sincronizados; rotinas com fundamentos e provas de conceitos para genérica classificação e recomendação de recursos	finalizar pacotes oficiais da linguagem Python; estação de monitoramento massivo; sistema de navegação semântica enriquecido com recomendação de recursos
dados	dados triplificados do Participa.br, do Cidade Democrática, do AA	revisar dados triplificados; triplificar dados do Facebook, Twitter e listas de email
ontologias e vocabulários	OPS, OPa, OPP, OCD, Ontologiaa, OBS e VBS iniciais	ontologias e vocabulários revisados
audiovisualização	versinus; prelúdio social; four hubs dance	músicas focando em algum dos participantes da rede; mais músicas sobre as redes do Facebook; mais músicas sobre as redes de email; rotinas para fazer animação abstrata sobre rede de interação e mixar com clipe do youtube; sonificação de dados semânticos e renderização de imagens sincronizadas

6
Conclusões e previsão

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 e com componentes principais típicas e estáveis. Um legado de dados ligados e abertos é conveniente para apresentar estes resultados às comunidades acadêmicas e interessadas nas aplicações, para as quais foram adiantadas ontologias, vocabulários, rotinas de conversão de dados relacionais em RDF e os dados em si.

Uma direção simples para concluir a pesquisa consiste em focar no documento *Complex Networks Gradus ad Parnassum*, que está planejado como uma apresentação das redes complexas através da entrega, para o leitor, de formas de observar e interagir com suas redes, beneficiando-se. Uma direção menos pedagógica, porém mais usual e simples, é explorar as estabilidades encontradas: até que número de agentes a distribuição dos participantes nos setores e a formação das componentes principais se mantém? Para quais redes? Como caracterizar a intermitência dos agentes enquanto a distribuição de grau é estável? Se o texto produzido pelos setores é diferente, em quais aspectos é igual e em quais se diferencia? Os resultados se mantém em ambas as línguas português e inglês?

Há, em alguns casos extremos, considerações na base da área, com implicações sobre a própria constituição das redes complexas (como na Seção **??**). Ao mesmo tempo, os métodos utilizados são potencialmente novos (como na Seção **??**). Há diversos trabalhos na bibliografia e, caso haja disponibilidade para visitar itens da literatura produzida, recomendamos, nesta mesma ordem, (**?**, **?**, **?**, **?**, **?**).

Agradecimentos

CNPq, PNUD/ONU, IEA/USP, Nexus/CNPq, SGPR, labMacambira.sf.net.