UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

RENATO FABBRI

Estabilidade topológica e diferenciação textual em redes de interação humana:

redes complexas para o participante e a física antropológica

São Carlos
2015

RENATO FABBRI

Estabilidade topológica e diferenciação textual em redes de interação humana: redes complexas para o participante e a física antropológica

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. *Estabilidade topológica e diferenciação textual em redes de interação humana: redes complexas para o participante e a física antropológica*. 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, fornece 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 acentuada 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 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: 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; melhor documentação e desenvolvimento do aparato em software, ontologias e dados.

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

Os primeiros estudos sobre redes de interação humana datam do século XIX. Já 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. (1) 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 (2) a ciências sociais (3) e física (4, 5), 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, (4, 5) com a qual algumas características foram reveladas. Por exemplo, a topologia das redes de interação humana exibe 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, acompanhando uma transição de fase histórica. (6) 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). (7)

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, e apresentam inovações. 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 de redes complexas. A Seção 1.1.1 faz observações pontuais sobre cada área secundária. A Seção 1.1.2 expõe a proliferação de ambiguidades e sinônimos no jargão deste trabalho. A Seção 2 é dedicada aos dados analisados. A Seção 3 contém os métodos usados para atingir os resultados, que

são explicitados na Seção 4. O cronograma de atividades e uma comparação entre afazeres planejados e finalizados estão na Seção 5. A monografia termina com as conclusões na Seção 6, seguida de agradecimentos e referências.

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, (1) 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 extensão das estruturas maiores e como exemplos de propriedades. 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 ambos os pontos.

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. As medidas costumam se referir a um vértice, aresta, rede ou comunidade.
- 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, por exemplo, pode ser frequentemente compreendida como redes complexas em sistemas sociais humanos. Uma consideração cuidadosa dos livros e artigos lidos para esta pesquisa está na Seção 5.

1.1.1 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 têm impacto sobre o que está feito, e sendo feito, e foram cursadas formalmente uma disciplina sobre cada uma para o doutorado (veja Seção 5). Seguem informações pontuais sobre cada área.

Os termos processamento de linguagem natural (PLN) e mineração de texto (MT) podem em geral ser 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 a passos firmes como diretrizes para governos, acadêmicos e sociedade civil.

1.1.2 Ambiguidades e sinônimos no jargão

Além de recente, a área de redes complexas conflui 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. (**?**, 1) 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 3.3.
- *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.
- Laço, *loop*, *selfloop*, *autoloop*, *buckle* são termos usados para designar uma aresta de um vértice para ele próprio.

Neste trabalho, muitas outras questões sobre a nomenclatura merecem exposição para evitar entendimentos errados. Por falta de espaço, esta discussão ficará de fora desta monografia, mas deverá constar em. (8)

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, (9) que consiste em mais de 20 mil listas de email e mais de 130 milhões de mensagens públicas. (10) 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. (3, 11)

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 (12), mas dezenas, talvez algumas centenas, foram observadas nos experimentos da Seção 4.4.2.
- Milhares de tweets (talvez alguns milhões), geralmente vinculados a alguma *hashtag*. Em especial, a rede de *retweets* de 22 mil *tweets* com a *hashtag* #arenaNETmundial, foi analisada em. (12)
- Mecanismos participativos como o Participa.br, Cidade Democrática e o AA. As redes de amizade e de interação do Participa.br foram analisadas em. (12)

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 (12) 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. (13)
- 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 relevantes para os resultados.
- 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 nesta pesquisa de considerações qualitativas e do cânone das ciências humanas.
- A Seção 3.8 delineia muito brevemente 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: cada unidade de tempo foi considerado um intervalo em que foram contabilizadas as atividades (e.g. mensagens de email). Também foram feitas algumas medidas de estatística circular, (14) conforme exposto a seguir.

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 (12), 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.

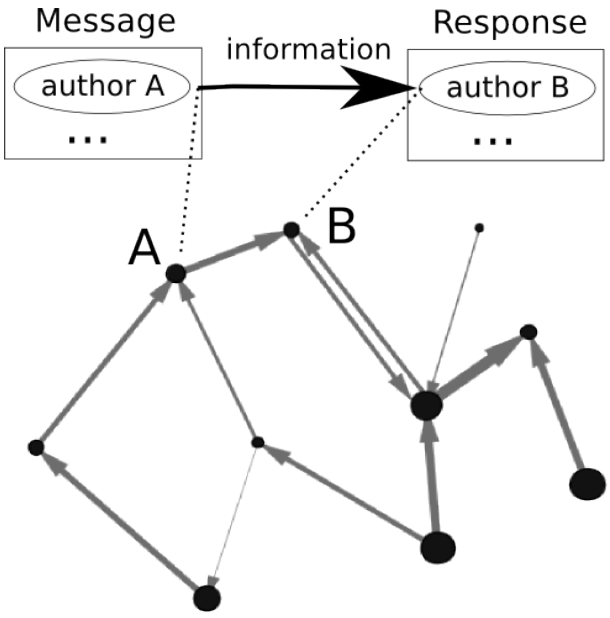
3.2 Formação das redes de interação

Redes de interação podem ser modeladas tanto com quanto sem peso, tanto dirigida quando não dirigida. (1,3,15,16) 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 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 dirigidas 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 1; a cada nova interação $A \rightarrow B$ é adicionado 1 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. (1, 3)

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



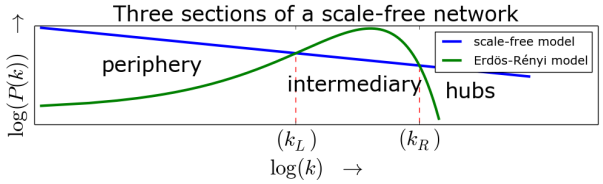
Fonte: Adaptada de FABBRI (12)

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 (13), 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).

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).



Fonte: Adaptada de FABBRI (12)

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. 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. Assim, podemos escrever que:

$$\sum_{x=k_i}^{k_j} \widetilde{P}(x) < \sum_{x=k_i}^{k_j} P(x) \Rightarrow \text{i é intermediário} \tag{3.3}$$

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 arestas, portanto mesma probabilidade p_e para existência de aresta, terá grau k com probabilidade $P(k)$ 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. 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:

$$\hat{P}(k^{way}) = \binom{N-1}{k^{way}} p_e^k (1-p_e)^{N-1-k^{way}} \tag{3.5}$$

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:

$$\begin{aligned} e_{\gamma,1} &= \{\ i \mid \overline{k}_{\gamma,L} \geq \overline{k}_{\gamma,i}\} \\ e_{\gamma,2} &= \{\ i \mid \overline{k}_{\gamma,L} < \overline{k}_{\gamma,i} \leq \overline{k}_{\gamma,R}\} \\ e_{\gamma,3} &= \{\ i \mid\ \ \overline{k}_{\gamma,i} < \overline{k}_{\gamma,R}\} \end{aligned} \tag{3.6}$$

onde $\overline{k}_{\gamma,i}$ é a medida γ no vértice i, convencionada:

$$\begin{aligned} \overline{k}_{1,i} &= k_i \\ \overline{k}_{2,i} &= k_i^{in} \\ \overline{k}_{3,i} &= k_i^{out} \\ \overline{k}_{4,i} &= \frac{s_i}{\overline{w}} \\ \overline{k}_{5,i} &= \frac{s_i^{in}}{\overline{w}} \\ \overline{k}_{6,i} &= \frac{s_i^{out}}{\overline{w}} \end{aligned} \tag{3.7}$$

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. Utilizando as Equações 3.6, estes critérios compostos C_δ, com δ inteiro no intervalo $1 \leq \delta < 6$ podem ser descritos como:

$$
\begin{aligned}
C_1 &= \{c_{1,\phi} = \{i \mid i \ \in e_{\gamma,\phi},\ \forall\, \gamma\}\} \\
C_2 &= \{c_{2,\phi} = \{i \mid \exists\ \gamma : i \in e_{\gamma,\phi}\}\} \\
C_3 &= \{c_{3,\phi} = \{i \mid i \ \in e_{\gamma,\phi'},\ \forall\, \gamma,\ \forall\, \phi' \geq \phi\}\} \\
C_4 &= \{c_{4,\phi} = \{i \mid i \ \in e_{\gamma,\phi'},\ \forall\, \gamma,\ \forall\, \phi' \leq \phi\}\} \\
C_5 &= \{c_{5,\phi} = \{i \mid i \ \in e_{\gamma,\phi'},\ \forall\, \gamma, \\
&\qquad\qquad \forall\ (\phi'+1)\%4 \leq (\phi+1)\%4\}\} \\
C_6 &= \{c_{6,\phi} = \{i \mid i \ \in e_{\gamma,\phi'},\ \forall\, \gamma, \\
&\qquad\qquad \forall\ (\phi'+1)\%4 \geq (\phi+1)\%4\}\}
\end{aligned}
\tag{3.8}
$$

No artigo (12), os critérios C_1, C_3 e C_5 foram chamados exclusivistas, os critérios C_2, C_4 e C_6 foram chamados inclusivistas, 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. (12))

3.4 Médias e variâncias nas Análises de Componentes Principais de cada sistema

A Análise de Componentes Principais (PCA é a sigla consagrada, do inglês Principal Component Analysis) é bastante estabelecida e bem documentada e foi usada para saber: 1) quais as medidas que contribuem para cada componente principal e em que proporção; 2) quanto da dispersão está concentrada em cada componente. 3) médias e desvios padrão destas quantidades para vários sistemas diferentes. Em geral, os diferentes sistemas eram evolução temporal um do outro.

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, $V'[j,k] = 100 * \frac{|V[j,k]|}{\sum_{j'} |V[j',k]|}$ dá a percentagem da componente principal k que 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 observam-se 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 (12), 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:

$$\begin{aligned}
\mu_{V'}[j,k] &= \frac{\sum_l^L V'[j,k,l]}{L} \\
\sigma_{V'}[j,k] &= \sqrt{\frac{(\mu_{V'} - V'[j,k,l])^2}{L}} \\
\mu_{D'}[k] &= \frac{\sum_l^L D'[k,l]}{L} \\
\sigma_{D'}[k] &= \sqrt{\frac{(\mu_{D'} - D'[k,l])^2}{L}}
\end{aligned} \tag{3.9}$$

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 redes foi estudada utilizando PCA (17) 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*). (1) Além disso, para apreender as simetrias das atividades dos participantes, as seguintes métricas foram introduzidas para cada 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:

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

onde $D_{n,n'} = sup_x[F_{1,n} - F_{2,n'}]$ (a maior diferença entre as duas cumulativas) e $c(\alpha)$ é tabelado para cada região crítica α (probabilidade da hipótese nula ser verdadeira).

São calculados $D_{n,n'}$, enquanto n e n' são dados. Todos os termos da Equação 3.10 são positivos e $c(\alpha)$ pode ser isolado:

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

Utilizamos $c'(\alpha)$ como distância entre pares de distribuições empíricas, o que é coerente com a teoria (18).

3.6 Audiovisualização de dados

Redes foram visualizadas com imagens, videos e engenhocas online para esta pesquisa (19–21). Redes também foram sonificadas, em especial como faceta sonora de animações abstratas. (9, 22–24) 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. (19)

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? 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 (25), leituras (26), escrita (27, 28), e contatos individuais com outros pesquisadores.

3.8 Web semântica

As estruturas sociais são muitas vezes ditadas por estruturas pré-concebidas, fruto de tradições e esforços especialistas. Para a formalização de conceitualizações, e associados formatos de dados apropriados para armazenamento compartilhamento e referência, foram adotadas as recomendações de dados ligados / web semântica da W3C. (29, 30) 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. (30) Os vocabulários são coleções de termos e metadados, como definição, e a recomendação é o uso do padrão SKOS. (31) A web semântica tem apresentado avanços: as inferências, por exemplo, têm se tornado mais ágeis e úteis, especialmente para buscas. Ao mesmo tempo, 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 identificada com um conceito SKOS é, por consequência, um indivíduo. Neste caso, quando 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. (32)

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 (12), 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 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 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 relacionado a outras características topológicas e textuais da rede e pode servir para uma tipologia das próprias redes.

- 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. (12)

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. (33) 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, menores 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 centralidade de 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 à prevalência de medidas topológicas **ou** textuais, mas a combinação de medidas de ambos os tipos é incidente.

Estes resultados permearam várias outras frentes de pesquisa e desenvolvimento tecnológico. (32, 34–37)

4.2 Criação da nuvem brasileira de dados participativos ligados

Iniciada para formalizar as redes e participantes, estabelecer *benchmarks* (valores de referência), e observar os aspectos mandatórios, e relativamente estáveis, das conceitualizações

sobre as estruturas sociais. 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 associados às 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. (27, 32, 34–36) Foi publicado no arXiv somente um artigo sobre a OPS. (36) Este escrito aguarda confluência com orientador para publicação em revista, potencialmente na revista PLOS ONE. Foram escritos também os produtos PNUD/ONU, publicados em instâncias governamentais e em repositórios públicos. (32, 34, 35) Foram publicados dados em RDF do Participa.br, Cidade Democrática e AA no Datahub.io. (38) Ontologias e vocabulários foram publicadas junto ao ministério do planejamento e em repositórios públicos (32). *Scripts* para síntese das ontologias e triplificação de dados estão também publicamente acessíveis e junto aos produtos PNUD da bibliografia.

4.2.1 Síntese de ontologias e vocabulários de estruturas sociais

Ontologias OWL feitas neste trabalho:

- OPS (Ontologia de Participação Social, fruto de diversos esforços da América Latina, principalmente do Brasil): nesta pesquisa, revisamos a ontologia e disponibilizamos a versão em uso por instâncias diferentes da academia, Estado e sociedade civil. (36)
- OPa (Ontologia do Participa.br): esta é uma ontologia feita para e a partir dos dados do Participa.br (Portal Federal de Participação Social, SG-PR).
- OPP (Ontologia de Portais Participativos): 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): uma pequena ontologia para o minimalista AA (Autorregulação Algorítmica), um software para registrar e compartilhar processos intelectuais como para pesquisa e arte. (32, 39, 40)
- 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 4.2.3 e utilizado também para a construção da OPa (acima).
- OBS (Ontologia da Biblioteca Social): uma coleção de ontologias, uma para cada conceito que precisasse, e uma para cada mecanismo ou instância de participação social

prevista no Decreto Presidencial nº 8.243, conhecido como decreto da PNPS ou da Política Nacional de Participação Social. Esta ontologia contou com entrevistas feitas diretamente para construí-la, e uma atividade especial 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. (32)

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

As ontologias e vocabulários são todas construídas através de scripts, com exceção da OPP, feita no Protegé. (32)

4.2.2 Obtenção de dados ligados a partir de dados relacionais participativos

Roteiros para conversão de dados relacionais em dados RDF enriquecidos semanticamente (32):

- Triplificação do Participa.br: dados originalmente em PostgreSQL. São usados, através de buscas SparQL, para auxiliar na construção da OPa.
- Triplificação dos dados do Cidade Democrática. Estes dados são utilizados para auxiliar na construção da OCD.
- Triplificação dos dados do AA: dados do AA encontrados em bancos de dados MySQL e MongoDB, e em *logs* de IRC. (32, 40) Esta foi a única triplificação feita depois da ontologia e não aproveitada para a construção da ontologia.

4.2.3 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 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. (32)

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. (9) Estão sendo feitos pacotes para organizar os numerosos *scripts* de triplificação de dados, construção de ontologias e vocabulários e mineração das estruturas. (41, 42) 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 (32). Muitas engenhocas foram criadas para gerar figuras, vídeos e inspecionar estruturas sociais de emails, Facebook, Twitter, Participa.br, AA, IRC e outras fontes. (21, 24, 27, 40) Outras engenhocas foram criadas para experimentações estéticas e informacionais. (9, 43–45)

4.4 Benefício, utilidade e formalismo

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 prova de conceito (35):

- 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 e a física antropológica

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 e hipóteses serão expostas nesta seção. 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, todas redes das quais faço parte. Foram confirmadas as hipóteses: de modificação das estruturas sociais para comportar a pesquisa, com suporte humano, financeiro e institucional; de modificação do tratamento da sociedade sobre o tema; de que seriam verificáveis estes resultados em minhas interações cotidianas. 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, três vezes. (46)

Em outra ocasião, percebi que uma característica não intuitiva: em praticamente qualquer rede de amizades com mais de 500 pessoas, dentre as 50 pessoas com a maior intermediação (*betweenness centrality*, mais participa de geodésicas) havia sempre pouquíssimas que constavam também dentre as 50 com maior closenness centrality (mais perto de todos os outros agentes). Selecionamos estes dois grupos em minha rede e em redes de parceiros que fazem

experimentos semelhantes. Cada pessoa enviou uma mensagem diferente, cada grupo de cada pessoa recebeu uma cópia desta mensagem. O grupo com a maior intermediação reagiu sempre calorosamente, repassava a mensagem, os membros até interagiam entre si, mesmo sem se conhecerem ou serem próximos. Os grupos de maior *closenness* nunca reagiam, membros saíam rapidamente da interface. A hipótese mais plausível que surgiu para explicar esta diferença de reação é a de que os membros 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 (*#arenaNETmundial*), 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 reflexão com as pessoas que foram ao evento, inclusive com os próprios comunicadores que constavam nas redes de *retweet*. (44) 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. (2, 47)

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.

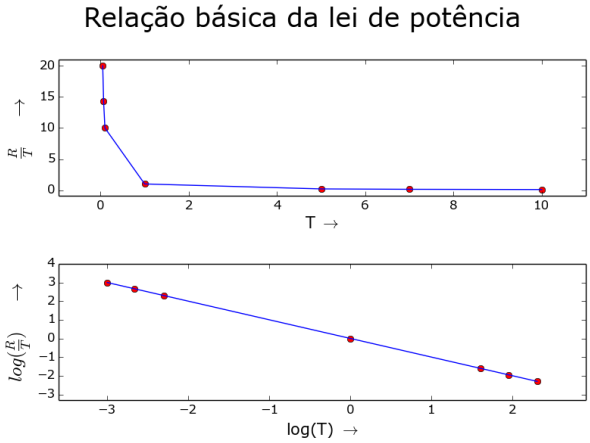
Estas diretrizes foram apreendidas em grande parte da tradição antropológica, e, portanto, configuram uma pesquisa com alguns aspectos "antropológicos". O termo "física antropológica" começou a ser usado no Brasil principalmente por acadêmicos (físicos, cientistas da computação, filósofos, antropólogos) 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 (7) e sociofísica. (48) Há resistência do meio científico, mas no geral o balanço aponta para uma pertinência do uso do termo para representar o que está sendo feito.

4.4.3 Entendimento sobre as estruturas sociais

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. (8) Um exemplo especial de fundamentação que parece não constar na literatura é a constatação de que a propriedade livre de escala apresenta uma distribuição uniforme de recursos com respeito à conectividade (e.g. grau ou força). Para apreender este fato, considere uma quantidade fixa R de recursos que será utilizada para a realização da rede em conectividade. Considere que, para cada quantidade de recursos T, são contadas $f = \frac{R}{T}$ partes de tamanho T, como na Figura 4.1.

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.



Fonte: Elaborada pelo autor.

Segue que $log(f) = -1log(T) + C$, com $C = log(R)$ uma constante arbitrária. Uma reta descreve a relação entre $log(f)$ e $log(T)$, como na Figura 4.1. Os recursos são alocados pelo sistema de forma uniforme, pois $T\frac{R}{T} = R = constante$.

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 - Cronograma e afazeres

Tabela 5.1 – Cronograma de atividades ao longo dos semestres, descritas na Seção 5. 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	[•]	[•]	[•]	[•]	[•]	[•]

Fonte: Elaborada pelo autor

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

1. Cumprimento dos créditos obrigatórios em disciplinas*.
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 ‖.

*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). No mestrado, fazia mais de 20 créditos na por semestre na 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.

†A literatura para o trabalho proposto é ampla e este aprofundamento tem sido constante.

‡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.

§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 4.4.3.

¶Conseguimos finalizar um artigo (12). Há ao menos mais dois em condições de publicação (36, 49). Além destes, há mais estes artigos no arXiv (27,33,37,50), todos referentes ao trabalho do doutorado. Foram publicados em revista os artigos sobre o AA (39) e análise quantitativa de pintura (51), ambos sem a colaboração do orientador.

‖Os experimentos de coleta e difusão de informação dispararam 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 à pesquisa, sobre o qual a Presidência da República se posicionou como beneficiária (27).

Tabela 5.2 – Relação de tarefas feitas e por fazer. Há literatura pronta e vários documentos escritos e em mãos para serem aprofundados. O mais urgente parece ser uma revisão e aprofundamento de estatística e física estatística, e confirmar os experimentos percolatórios contínuos (veja Seção 4.4.2).

	feito	**por fazer**
escrita	artigo de estabilidade em redes de interação humana (12); artigo sobre a Ontologia de Participação Social (36); ensaio descrevendo simbiose com PNUD/ONU e SG-PR (27); artigo com descrição psicofísica da música no áudio digital (49); 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 (32, 34, 35); artigo sobre AA (39); versões iniciais e rascunhos dos artigos sobre física antropológica (28), sobre votação contínua por aprovação e participação (50), sobre diferenças da produção textual nos setores de Erdös (33), sobre visualização de redes de interação em evolução temporal (37), sobre audiovisualização de redes de interação em evolução temporal (37), sobre performance audiovisual via controle coletivo de código e projeção ao vivo (52)	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" (42); 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 (40)
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 (4, 16, 53, 54)	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
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

Fonte: Elaborada pelo autor.

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 da Seção 4.4.2 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. Além disso, foram consideradas as conceitualizações sobre as estruturas sociais, que são também relativamente estáveis, estruturas em rede e bastante caracterizáveis por palavras. Um legado de tais conceitualizações formalizadas e associadas a dados ligados e abertos é conveniente para *benchmarks* e 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.

Há, em alguns casos extremos, considerações na base da área, com implicações sobre a própria constituição das redes complexas (como na Seção 4.4.3). Ao mesmo tempo, os métodos utilizados são potencialmente novos (como na Seção 3.3). Há diversos trabalhos na bibliografia e, caso haja disponibilidade para visitar itens da literatura produzida, recomendamos, nesta mesma ordem, (9, 12, 23, 27, 32). Dois exemplos de trabalhos de terceiros com influência direta do material nesta monografia, são a tese de doutorado defendida ano passado; e o relatório de consultoria prestada por professores da UnB ao PNUD junto à SG-PR. (55,56)

A frente audiovisual possibilitou arte e inspeção dos dados, orientou a pesquisa e facilitou a comunicação com especialistas e não especialistas. Foram feitas músicas, animações abstratas, vídeos e imagens de estruturas sociais, (22, 23, 37, 57) muitas vezes com outras pessoas e com as redes delas, como explicitado na Seção 4.4.2. De uma forma geral, o audiovisual tornou o pensamento sobre as redes de interação humana mais familiar, acessível e convidativo.

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. (8) 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 mantêm em ambas as línguas português e inglês?

O orientador solicitou ênfase do conteúdo do artigo de estabilidade temporal (12) nesta monografia, o que me esforcei por fazer. Os complementos principais ao conteúdo aqui apresentado são: a seção de resultados do corpo do artigo, e o documento de informações de suporte. (57)

Agradecimentos

Autores agradecem o apoio financeiro concedido pelo CNPq (140860/2013-4, projeto 870336/1997-5), Programa das Nações Unidas para o Desenvolvimento (contrato 2013/000566, projeto BRA/12/018) e FAPESP; a prestatividade do corpo do IFSC, da CPG e dos funcionários; o espaço concedido pelo IEASC/USP para algumas reuniões e visitas de parceiros de pesquisa; o suporte intelectual e o interesse dos membros do Nexos-Sudeste (Grupo de pesquisa/CNPq (25)); ao suporte institucional e intelectual da Secretaria Geral da Presidência da República; ao suporte intelectual e tecnológico do labMacambira.sf.net e todas as comunidades de software e cultura livre direta e indiretamente relacionadas a este trabalho.

REFERÊNCIAS

1 NEWMAN, M. *Networks*: an introduction. Oxford: Oxford University Press, 2010.

2 LATOUR, B. Reassembling the social. an introduction to actor-network-theory. *Journal of Economic Sociology*, v. 14, n. 2, p. 73-87, 2013.

3 BIRD, C. et al. *Mining email social networks*. 2006. P. 137-143. Disponível em: <http://macbeth.cs.ucdavis.edu/msr06.pdf >. Acesso em: 16 jun. 2015.

4VÁZQUEZ, A. et al. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, v. 73, n. 3, p. 036127, 2006.

5 BRIAN BALL, B.; NEWMAN, M. E. J. *Friendship networks and social status*. 2012. Disponível em: < http://arxiv.org/pdf/1205.6822v1.pdf >. Acesso em: 16 jun. 2015.

6 PETLAND, A. *Reinventing society in the wake of big data*. 2012. Disponível em: < http://edge.org/conversation/reinventing-society-in-the-wake-of-big-data>. Acesso em: 16 jun. 2015.

7 PETLAND, A. *Social physics*: how good ideas spread-the lessons from a new science. New York: Penguin Press, 2014.

8 FABBRI, R. *Complex networks gradus ad parnassum*. 2015. Disponível em: <https://github.com/ttm/gradus/raw/master/article.pdf>. Acesso em 16 jun. 2015.

9 FABBRI, R. *Python package to observe time stability in the gmane database*, 2015. Disponível em: <https://pypi.python.org/pypi/gmane>. Acesso em: 16 jun. 2015.

10 WIKIPEDIA. *Gmane*. 2013. Disponível em: < https://en.wikipedia.org/wiki/Gmane> Acesso em: 27out. 2013.

11 MAREK-SPARTZ, K.; CHESLEY, P.; SANDE, H. *Construction of the gmane corpus for examining the difusion of lexical innovations*. 2012. Disponível em: <http://people.lis.illinois.edu/~jdiesner/calls/papers_all/WON__2012_Spartz_Chesley_Sande_Gmane_Corpus.pdf>. Acesso em: 15 jun. 2014.

12 FABBRI, R. *Time stability in human interaction networks*. 2015. Disponível em: < http://arxiv.org/pdf/1310.7769v6.pdf >. Acesso em: 15 jun. 2014

13 JACKSON, M. O. *Social and economic networks*: models and analysis. 2013. Disponível em: < http://online.stanford.edu/course/social-and-economic-networks-models-and-analysis >. Acesso em: 15 jun. 2014

14 WIKIPEDIA. *Directional statistics*. 2015. Disponível em: <https://en.wikipedia.org/wiki/Directional_statistics>. Acesso em: 16 jun. 2015.

15 LEICHT, E. A.; NEWMAN, M. E. J. Community structure in directed networks. *Physical Review Letters*, v. 100, n. 11, 118703, 2008.

16 NEWMAN, M. E. J. *Community detection and graph partitioning*. 2013. Disponível em: <http://arxiv.org/pdf/1305.4974v1.pdf>. Acesso em: 15 jun. 2014

17 JOLLIE, I. *Principal component analysis.* New York: Wiley Online Library, 2005.

18 WIKIPEDIA. *Kolmogorov-smirnov test*. 2015. Disponível em: <https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test>. Acesso em 12 jun. 2015.

19 FABBRI, R. *Video visualizations of email interaction network evolution*. Disponível em: <https://www.youtube.com/watch?v=AkmtkCHmQp4&list=PLf_EtaMqu3jUYO_XfJdqQELdbFnpqYEfb>. Acesso em 15 jun. 2015.

20 FABBRI, R. *Image gallery of email interaction networks*. 2013. Disponível em: <http://hera.ethymos.com.br:1080/redes/python/autoRede/gmane.linux.audio.devel_3000-4200-280/>. Acesso em: 15 jun. 2014

21 FABBRI, R. *Online gadget for making email interaction network images, gml les and measurements*. 2013. Disponível em: <http://hera.ethymos.com.br:1080/redes/python/autoRede/escolheRedes.php>. Acesso em: 15 jun. 2014

22 FABBRI, R. *Prelúdio social* (audiovisualização de rede social). 2015. Disponível em: <https://www.youtube.com/watch?v=9uIiGQBuWYo>. Acesso em: 15 jun. 2014

23 FABBRI, R. *Four hubs dance* (audiovisualização de rede social). 2015. Disponível em: https: <//www.youtube.com/watch?v=1EOnEVmqmmc>. Acesso em: 15 jun. 2014

24 FABBRI, R. *Python package for social data scraping, analysis and art*. 2015. Disponível em: <https://pypi.python.org/pypi/social>. Acesso em: 15 jun. 2014

25 GRUPO nexos: teoria crítica e pesquisa interdisciplinar. Disponível em: <http://dgp.cnpq.br/dgp/espelhogrupo/5624425913774111>. Acesso em: 5 jun. 2015.

26 CANEVACCI, M. *Antropologia, psicanálise, comunicação*. 2012. Disponível em: <https://goo.gl/rZoR9D>. Acesso em 15 jun. 2015.

27 FABBRI, R. *Ensaio sobre o auto-aproveitamento*: um relato de investidas naturais na participação social. 2014. Disponível em: < http://arxiv.org/pdf/1412.6868v2.pdf>. Acesso em 15 jun. 2015.

28 FABBRI, R. *What are you and i?* [anthropological physics fundamentals]. 2015. Disponível em: <https://www.academia.edu/10356773/What_are_you_and_I_anthropological_physics_fundamentals_>. Acesso em 15 jun. 2015.

29 WORLD WIDE WEB CONSORTIUM. et al. Rdf 1.1 concepts and abstract syntax. 2014. Disponivel em: <http://www.w3.org/TR/rdf11-concepts/>. Acesso em 15 jun. 2015.

30 WORLD WIDE WEB CONSORTIUM. et al. *Owl 2 web ontology language document overview*. 2012. Disponível em: <http://www.w3.org/TR/owl2-overview/>. Acesso em 15 jun. 2015.

31 MILES, A.; BECHHOFER, S. *Skos simple knowledge organization system reference.W3C recommendation, 18:W3C*. 2009. Disponível em: <http://www.w3.org/TR/skos-reference/>. Acesso em 15 jun. 2015.

32 FABBRI, R. *Proposta de regras de extração de conteúdos da api do portal e suas ferramentas para alimentação de eventual/hipotética base/nuvem de conhecimento de participação social*. 2014. Disponível em: <https://github.com/ttm/pnud5/blob/master/latex/produto.pdf?raw=true>. Acesso em: 10 jun. 2015.

33 FABBRI, R. *A connective diferentiation of textual production in interaction networks*. 2013. Disponível em: <http://arxiv.org/abs/1412.7309>. Acesso em: 10 jun. 2015.

34 FABBRI, R. *Ferramentas assistidas de categorização de conteúdo*: com processamento de linguagem natural e de redes complexas, adaptadas para o ambiente do portal federal de participação social (Participa.br). 2014. Disponível em: <https://github.com/ttm/pnud3/blob/master/latex/produto.pdf?raw=true>. Acesso em: 10 jun. 2015.

35 FABBRI, R. *Adaptações e incrementos para a interface do portal federal de participação social e suas ferramentas*. 2014. Disponível em: <https://github.com/ttm/pnud4/blob/master/latex/produto.pdf?raw=true>. Acesso em: 10 jun. 2015

36 FABBRI, R. et al. *Social participation ontology*: community documentation, enhancements and use examples. Disponível em: <http://arxiv.org/pdf/1501.02662v2.pdf>. Acesso em: 10 jun. 2015.

37 FABBRI, R. *Versinus*: a visualization method for graphs in evolution. 2013. Disponível em: <http://arxiv.org/abs/1412.7311>. Acesso em 12 jun. 2015.

38 FABBRI, R. *Brazilian social participation data from Participa.br*, Cidade Democrática and AA, in XML/RDF and Turtle/RDF. 2014. Disponível em: <http://datahub.io/organization/socialparticipation>. Acesso em: 12 jun. 2015.

39 FABBRI, R. et al. The algorithmic autoregulation software development methodology. *Revista Eletrônica de Sistemas de Informação*, v. 13, n. 2, 2014. doi: 10.5329/RESI.

40 FABBRI, R. *The algorithmic-autoregulation essay*: a collective and natural focus on self-transparency. Disponível em: <https://github.com/ttm/ensaaio/raw/master/ensaio.pdf>. Acesso em: 12 jun. 2015.

41 FABBRI, R. *Python package with routines for analysis and synthesis of RDF social* participation data. 2015. Disponível em: <https://pypi.python.org/pypi/participation>. Acesso em: 12 jun. 2015.

42 FABBRI, R. *Python package for anthropological physics and social harnessing*. 2015. Disponível em: <https://pypi.python.org/pypi/percolation>. Acesso em: 24 maio 2015.

43 MMISSA: Monitoramento massivo e interativo da sociedade pela sociedade para aproveitamento. Disponível em: <http://mmissa.meteor.com>. Acesso em: 16 dez. 2014

44 TELÕES de streaming de estruturas sociais para o #ocupagov. Disponível em: <http://ocupagov.meteor.com>. Acesso em: 16 dez. 2014.

45 MÚSICA social e govern art. Disponível em: <http://mm.meteor.com>. Acesso em: 16 dez. 2014

46 RECEITAS para percolação da própria rede. Disponível em: <https://dl.dropboxusercontent.com/u/22209842/doc/mit/difusao.pdf>. Acesso em: 15 jun. 2014.

47 FABBRI, R. Nuvens cognitivas e a unificação da espécie humana. *CyberiuN Revista Eletrônica*. 2013. Disponível em: <http://wiki.nosdigitais.teia.org.br/Cyberiun>. Acesso em: 12 jun. 2015.

48 GALAM, S. *Sociophysic*s: a review of galam models. *International Journal of Modern Physics C*, v. 19, n. 03, p. 409-440, 2008.

49 FABBRI, R. et al. *Psychophysics of musical elements in the discrete-time representation of sound*. Disponível em: <http://arxiv.org/pdf/1412.6853v1.pdf>. Acesso em: 12 jun. 2015.

50 FABBRI, R.; POPPI, R. *Continuous voting by approval and participation*. Disponível em: <http://arxiv.org/abs/1505.06640>. Acesso em: 13 jun. 2015.

51 VIEIRA, V. et al. A quantitative approach to painting styles. *Physica A*: statistical mechanics and its applications, v. 417, p. 110-129, 2015.

52 VIEIRA, V. et al. Vivace: a collaborative live coding language. Disponível em: <http://arxiv.org/abs/1502.01312>. Acesso em: 12 jun. 2013.

53 CLAUSET, A.; SHALIZI,C. R.; NEWMAN, M. E. J. Power-law distributions in empirical data. *SIAM Review*, v. 51, n. 4, p.661-703, 2009.

54 PALLA, G. et al. Quantifying social group evolution. *Nature*, v. 446, n. 7136, p.664-667, 2007.

55 VIEGAS. C. *Línguas em rede: para o fortalecimento da língua e da cultura Kokama*. 2015. Disponível em: <http://repositorio.unb.br/handle/10482/17521>. Acesso em: 13 jun. 2015.

56 CRUZ, F.; MEIRELES, P. *Extensão de ontologia para o Portal de Participação Social, Conferências Nacionais e Conselhos Nacionais, e articulação entre os dados desses instrumentos*. https://github.com/ttm/tese/blob/master/bib/sgpr/pnud_fwcruz_pmeirelles_ produto06_vfinal.docx?raw=true.

57 FABBRI, R. [*Supporting information*] time stability in human interaction networks. Disponível em: <http://arxiv.org/src/1310.7769v6/anc/supportingInformation.pdf>. Acesso em: 12 jun. 2015.

ANEXO

Time stability in human interaction networks

Renato Fabbri[1, a)]
São Carlos Institute of Physics, University of São Paulo (IFSC/USP)

(Dated: 13 June 2015)

In this study, we demonstrate a remarkably stable activity in human interaction networks. The activity along time and topology evolution were investigated in four e-mail lists by considering window sizes from 50 to 10,000 messages, which were made to slide and generate snapshots of the network in a timeline. Furthermore, the activity in terms of seconds, minutes, hours, days and months, is practically the same for all lists. The activity of participants followed the expected scale-free behavior, thus allowing us to establish three classes of vertices by comparing with the Erdös-Rényi model, namely hubs, intermediary and peripheral vertices. The relative size of these three sectors did not vary with time and was essentially the same for all e-mail lists. Typically, 3-12% of the vertices are hubs, 15-45% are intermediary and the remainder are peripheral vertices. The metrics that contribute most to the dispersion of participants in the topological measures space were centrality measurements (degree, strength and betweenness), followed by symmetry-related metrics and then clustering coefficient. Similar results for the distribution of participants in the three categories and for the relative importance of the topological metrics were obtained for 12 additional networks from Facebook, Twitter and Participa.br. Consistent with expectations driven from the literature, these properties may be general for human interaction networks, which has important implications in establishing a typology based on objective, quantitative criteria.

PACS numbers: 89.75.Fb,05.65.+b,89.65.-s
Keywords: complex networks, social network analysis, pattern recognition, statistics, anthropological physics

'The reason for the persistent plausibility of the typological approach, however, is not a static biological one, but just the opposite: dynamic and social.' - *Adorno et al, 1969, p. 747*

I. INTRODUCTION

Studies on human interaction networks have started long before modern computers, dating back to the nineteenth century, while the foundation of social network analysis is generally attributed to the psychiatrist Jacob Moreno in mid twentieth century[1]. With the increasing availability of data related to human interactions, research on these networks has grown continuously. Contributions can now be found in a variety of fields, from social sciences and humanities[2] to computer science[3] and physics[4,5], given the multidisciplinary nature of the topic. One of the approaches from an exact science perspective is to represent interaction networks as complex networks[4,5], with which several features of human interaction have been revealed. For example, the topology of human interaction networks exhibits a scale-free trace, which points to the existence of a small number of highly connected hubs and a large number of poorly connected nodes. The dynamics of complex networks representing human interaction has also been addressed [6,7], but only to a limited extent, since research is normally focused on a particular metric or task, such as accessibility or community detection[8,9].

In this paper we analyze the evolution of human interaction networks, by considering interaction in email lists as their representative. Using a timeline of activity snapshots with a constant number of contiguous messages in email lists, we found a remarkable stability for several of the network properties. Because these properties were shared by networks from Twitter, Facebook and Participa.br, and are consistent with the literature, we advocate that some of the conclusions might be valid for more general classes of interaction networks. In particular, this allows us to discuss typologies in the context of such networks, in an attempt to bridge the gap between approaches based solely on data analysis (i.e. from a hard sciences perspective) and those relevant to the social sciences. This is important insofar as typologies are the canon of scientific literature for classification of human agents[10].

The paper is organized as follows. Section I A describes related work, while details of the data and methods of analysis are given in Section II and Section III. Section IV brings the results and discussion, leading to Section V for conclusions. Details of how to access the data are given in the Appendix, while subsidiary results from the email lists and of networks from Twitter, Facebook and Participa.br are given in the Supporting Information.

A. Related work

Research on network evolution often considers solely network growth, in which there is a monotonic increase

a) http://ifsc.usp.br/~fabbri/; Electronic mail: fabbri@usp.br

in the number of events considered[6]. Exceptions are reported in this section, with emphasis on those more closely related to the present article.

Network types have been discussed with regard to the number of participants, intermittence of their activity and network longevity[6]. Two topologically different networks emerged from human interaction networks, depending on the frequency of interactions, which can either be a generalized power law or an exponential connectivity distribution[11]. In email list networks, scale-free properties were reported with $\alpha = 1$[3] (as are web browsing and library loans[4]), and different linguistic traces were related to weak and strong ties[12].

Unreciprocated edges often exceed 50% in the networks analyzed, which matches empirical evidence from the literature[7] and motivated the inclusion of symmetry metrics in our analysis. No correlation of topological characteristics and geographical coordinates was found[13], therefore geographical positions were not considered in our study. Gender related behavior in mobile phone datasets was indeed reported[14], but this was not considered in the present article because email messages and addresses have no gender related metadata[15].

II. DATA DESCRIPTION: EMAIL LISTS AND MESSAGES

Email list messages were obtained from the GMANE email archive[15], which consists of more than 20,000 email lists and more than 130,000,000 messages[16]. These lists cover a variety of topics, mostly technology-related. The archive can be described as a corpus with metadata of its messages, including sent time, place, sender name, and sender email address. The GMANE usage in scientific research is reported in studies of isolated lists and of lexical innovations[3,12].

We analyzed many email lists (and data from Twitter, Facebook and Participa.br) but selected only four in order to make a thorough analysis, from which general properties can be inferred. These lists, selected as representing both a diverse set and ordinary lists, are:

- Linux Audio Users list[17], with participants holding hybrid artistic and technological interests, from different countries. English is the language used the most. Abbreviated as LAU from now on.

- Linux Audio Developers list[18], with participants from different countries, and English is the language used the most. A more technical and less active version of LAU. Abbreviated LAD from now on.

- Development list for the standard C++ library[19], with computer programmers from different countries. English is the language used the most. Abbreviated as CPP from now on.

TABLE I. Columns $date_1$ and $date_M$ have dates of first and last messages from the 20,000 messages considered in each email list. N is the number of participants (number of different email addresses). Γ is the number of threads (count of messages without antecedent). $\overline{M}$ is the number of messages missing in the 20,000 collection, $100\frac{23}{20000} = 0.115$ percent in the worst case. A relation holds for all these lists: within a same number of messages, as the number of participants increases, the number of threads decreases.

list	$date_1$	$date_M$	N	Γ	$\overline{M}$
LAU	2003-06-29	2005-07-23	1181	3372	5
LAD	2003-06-30	2009-10-07	1268	3109	4
MET	2005-08-01	2008-03-07	492	4607	23
CPP	2002-03-12	2009-08-25	1052	4506	7

- List of the MetaReciclagem project[20], with Brazilian activists holding digital culture interests. Portuguese is the most used language, although Spanish and English are also incident. Abbreviated MET from now on.

The first 20,000 messages of each list were considered, with total timespan, authors, threads and missing messages indicated in Table I. In subsidiary experiments we considered 140 additional email lists, also retrieved from the GMANE public database, to analyze the interdependence between the number of participants and the number of threads. Furthermore, we used 12 additional networks from Facebook (8), Twitter (2) and Participa.br (2) to grasp the generality of the results driven from email lists.

III. CHARACTERIZATION METHODS

The email lists and the networks generated from them were characterized using five procedures, namely: 1) statistics of activity along time, in scales from seconds to years; 2) sectioning of the networks in hubs, intermediary and peripheral vertices; 3) dispersion of basic topological metrics; 4) iterative visualization and data inspection. Each of these procedures are described below.

A. Time activity statistics

Messages were counted along time with respect to seconds, minutes, hours, days of the week, days of the month, and months of the year. This resulted in histograms from which patterns could be drawn. The ratio $\frac{b_h}{b_l}$ between the highest and lowest incidences on the histograms served as a hint of how close the observed distribution is to a uniform distribution.

The average and the dispersion were taken using circular statistics, in which each *measurement* (data point) is represented as a complex number with modulus equal to one, $z = e^{i\theta} = \cos(\theta) + i\sin(\theta)$, where $\theta = measurement\frac{2\pi}{period}$. The moments m_n, lengths of

moments R_n, mean angle θ_μ, and rescaled mean angle θ'_μ are defined as:

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

θ'_μ is used as the measure of location. Dispersion is measured using the circular variance $Var(z)$, the circular standard deviation $S(z)$, and the circular dispersion $\delta(z)$:

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

As expected, a positive correlation was found in all $Var(z)$, $S(z)$ and $\delta(z)$ dispersion measures, as can be noticed in Section I A of the Supporting Information, and $\delta(z)$ was preferred in the discussion of results.

B. Interaction networks

Interaction networks can be modeled both weighted or unweighted, both directed or undirected[3,21,22]. Networks in this article are directed and weighted, the more informative of trivial possibilities, i.e. we did not investigate directed unweighted, undirected weighted, and undirected unweighted representations of the interaction networks. The networks were obtained as follows: a direct response from participant B to a message from participant A yields an edge from A to B, as information went from A to B. The reasoning is: if B wrote a response to a message from A, he/she read what A wrote and formulated a response, so B assimilated information from A, thus $A \to B$. Inverting edge direction yields the status network: B read the message and considered what A wrote worth responding, giving status to A, thus $B \to A$. This article uses the information network as described above and depicted in Figure 1. Edges in both directions are allowed. Each time an interaction occurs, one is added to the edge weight. Selfloops were regarded as non-informative and discarded. These human social interaction networks are reported in the literature as exhibiting scale-free and small world properties, as expected for (some) social networks[1,3].

Edges can be created from all antecedent message authors on the message-response thread to each message author. We only linked the immediate antecedent to the new message author, both for simplicity and for the valid objection that in adding two edges, $x \to y$ and $y \to z$, there is also a weaker connection between x and z. Potential interpretations for this weaker connection are: double length, half weight or with one more "obstacles". This suggests the adequacy of centrality measurements to account for the connectivity with all nodes, such as betweenness centrality, eigenvector centrality and accessibility[8,23].

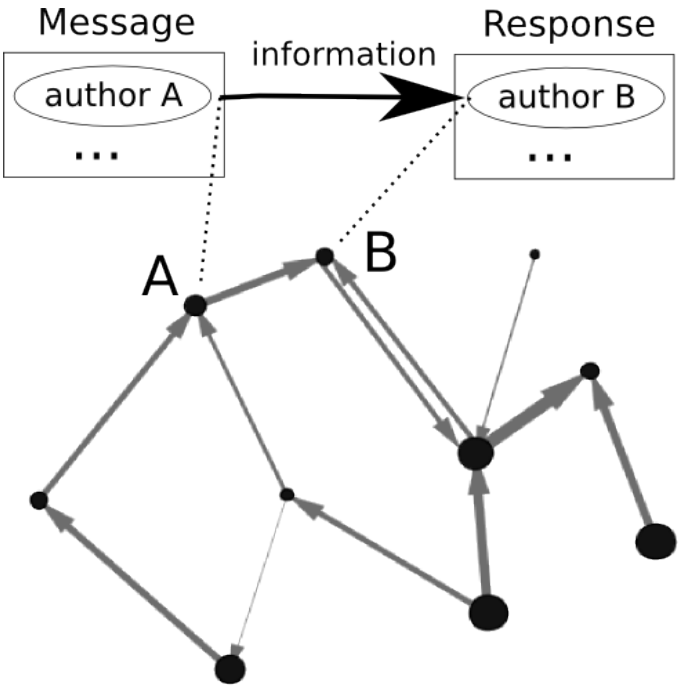


FIG. 1. Formation of interaction network from email messages. Each vertex represents a participant. A reply message from participant B to a message from participant A is regarded as evidence that B received information from A and yields a directed edge. Multiple messages add "weight" to a directed edge. Further details are given in Section III B.

C. Erdös sectioning

In a scale-free network, the peripheral, intermediary and hubs sectors can be derived from a comparison with an Erdös-Rényi network with the same number of edges and vertices[24], as depicted in Figure 2. We shall refer to this procedure as *Erdös sectioning*, with the resulting sectors being referred to as *Erdös sectors* or *primitive sectors*.

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

$$\widetilde{P}(k) < P(k) \Rightarrow \text{k is intermediary degree} \tag{3}$$

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

$$P(k) = \binom{2(N-1)}{k} p_e^k (1-p_e)^{2(N-1)-k} \tag{4}$$

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

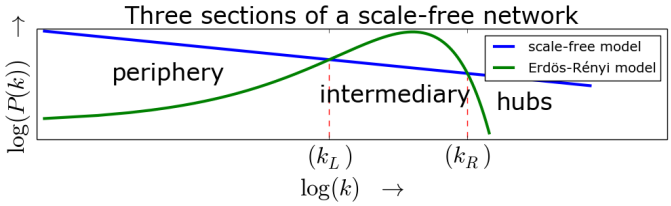


FIG. 2. Degree distributions of scale-free and Erdös-Rényi ideal networks. The latter has more intermediary vertices, while the former has more peripheral and hub vertices. The sector borders are defined by the two intersections k_L and k_R of the connectivity distributions. Characteristic degrees are in the compact intervals: $[0, k_L]$, $(k_L, k_R]$, $(k_R, k_{max}]$ for the Erdös sectors (periphery, intermediary and hubs).

To ensure statistical validity of the histograms, bins can be chosen to contain at least η vertices of the real network. Thus, each bin, starting at degree k_i, spans $\Delta_i = [k_i, k_j)$ degree values, where j is the smallest integer with which there are at least η vertices with degree larger than or equal k_i, and less than k_j. This changes equation 3 to:

$$\sum_{x=k_i}^{k_j} \widetilde{P}(x) < \sum_{x=k_i}^{k_j} P(x) \Rightarrow \text{i is intermediary} \tag{5}$$

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

$$\hat{P}(k^{way}) = \binom{N-1}{k^{way}} p_e^k (1-p_e)^{N-1-k^{way}} \tag{6}$$

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

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

$$\begin{aligned} e_{\gamma,1} &= \{\, i \mid \overline{k}_{\gamma,L} \ge \overline{k}_{\gamma,i}\} \\ e_{\gamma,2} &= \{\, i \mid \overline{k}_{\gamma,L} < \overline{k}_{\gamma,i} \le \overline{k}_{\gamma,R}\} \\ e_{\gamma,3} &= \{\, i \mid \quad\quad \overline{k}_{\gamma,i} < \overline{k}_{\gamma,R}\} \end{aligned} \tag{7}$$

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

$$\begin{aligned} \overline{k}_{1,i} &= k_i \\ \overline{k}_{2,i} &= k_i^{in} \\ \overline{k}_{3,i} &= k_i^{out} \\ \overline{k}_{4,i} &= \frac{s_i}{\overline{w}} \\ \overline{k}_{5,i} &= \frac{s_i^{in}}{\overline{w}} \\ \overline{k}_{6,i} &= \frac{s_i^{out}}{\overline{w}} \end{aligned} \tag{8}$$

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

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

- Exclusivist criterion C_1: vertices are only classified if the class is the same according to all metrics. In this case, vertices classified (usually) do not reach 100%, which is indicated by a black line in Figure 3.
- Inclusivist criterion C_2: a vertex has the class given by any of the metrics. Therefore, a vertex may belong to more than one class, and the total number of members may exceed 100%, which is indicated by a black line in Figure 3.

- Exclusivist cascade C_3: vertices are only classified as hubs if they are hubs according to all metrics. Intermediary are the vertices classified either as intermediary or hubs with respect to all metrics. The remaining vertices are regarded as peripheral.
- Inclusivist cascade C_4: vertices are hubs if they are classified as so according to any of the metrics. The remaining vertices are classified as intermediary if they belong to this category for any of the metrics. Peripheral vertices will then be those which were not classified as hub or intermediary with any of the metrics.
- Exclusivist externals C_5: vertices are only hubs if they are classified as such according to all the metrics. The remaining vertices are classified as peripheral if they fall into the periphery or hub classes by any metric. The rest of the nodes are classified as intermediary.
- Inclusivist externals C_6: hubs are vertices classified as hubs according to any metric. The remaining vertices will be peripheral if they are classified as such according to any metric. The rest of the vertices will be intermediary vertices.

Using equations 7, these compound criteria C_δ, with δ integer in the interval $1 \leq \delta < 6$ can be described as:

$$\begin{aligned}
C_1 &= \{c_{1,\phi} = \{i \mid i \in e_{\gamma,\phi},\ \forall\, \gamma\}\} \\
C_2 &= \{c_{2,\phi} = \{i \mid \exists\ \gamma : i \in e_{\gamma,\phi}\}\} \\
C_3 &= \{c_{3,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\, \gamma,\ \forall\, \phi' \geq \phi\}\} \\
C_4 &= \{c_{4,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\, \gamma,\ \forall\, \phi' \leq \phi\}\} \\
C_5 &= \{c_{5,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\, \gamma, \\
&\qquad \forall\, (\phi'+1)\%4 \leq (\phi+1)\%4\}\} \\
C_6 &= \{c_{6,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\, \gamma, \\
&\qquad \forall\, (\phi'+1)\%4 \geq (\phi+1)\%4\}\}
\end{aligned} \tag{9}$$

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

D. Topological metrics for Principal Component Analysis

The topology of the networks was studied using Principal Component Analysis (PCA[25]) with a small selection of the most basic and fundamental measurements for each vertex, as follows:

- Degree k_i: number of edges linked to vertex i.
- In-degree k_i^{in}: number of edges ending at vertex i.
- Out-degree k_i^{out}: number of edges departing from vertex i.
- Strength s: sum of weights of all edges linked to vertex i.
- In-strength s_i^{in}: sum of weights of all edges ending at vertex i.
- Out-strength s_i^{out}: sum of weights of all edges departing from vertex i.
- Clustering coefficient cc_i: fraction of pairs of neighbors of i that are linked. The standard clustering coefficient for undirected graphs was used.
- Betweenness centrality bt_i: fraction of geodesics that contain vertex i. The betweenness centrality index considered directions and weight, as specified in[26].

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

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

E. Evolution and visualization of the networks

The evolution of the networks was observed within a fixed number of messages, which we refer to as the window size ws. This same number of contiguous messages ws was considered with different shifts in the message timeline to obtain snapshots. Each snapshot was used both to perform the Erdös sectioning and to apply PCA for the topological metrics. The ws used were 50, 100,

TABLE II. The rescaled circular mean θ'_μ and the circular dispersion $\delta(z)$ described in Section III A. This typical table was made using all LAD list messages, and the results are the same for other lists, as shown in Section I A of the Supporting Information. Most uniform distribution of activity was found in seconds and minutes, where the mean has little meaning. Hours of the day exhibited the most concentrated activity (lowest $\delta(z)$), with mean between 14h and 15h ($\theta' = -9.61$). Weekdays, month days and months have mean near zero (i.e. near the beginning of the week, month and year) and high dispersion.

	θ'_μ	$\delta(z)$
seconds	−//−	9070.17
minutes	−//−	205489.40
hours	-9.61	4.36
weekdays	-0.03	29.28
month days	-2.65	2657.77
months	-0.56	44.00

200, 400, 500, 800, 1000, 2000, 2500, 5000 and 10000. Within a same ws, the number of vertices and edges vary in time, as do other network characteristics, which is exhibited in Section II of the Supporting Information.

Networks were visualized with animations, image galleries and online gadgets developed specifically for this research[27–29]. Such visualizations were crucial to guide research into the most important features of network evolution. Furthermore, the size of the three Erdös sectors could be visualized in a timeline fashion. Visualization of network structure was especially useful in the inspection of data and derived structures from the email lists.

IV. RESULTS AND DISCUSSION

A. Activity along time

The activity along time, in terms of seconds, minutes, hours, days and months, is practically the same for all lists. Histograms in each time scale were calculated as were circular average values and their dispersion. We chose to give detailed values in Table II-VI because these numbers can actually be used for characterizing nodes (participants) in other networks, and networks themselves, as they are independent of the network under analysis. For example, they may serve for identification of outliers in a community.

In the scale of seconds and minutes, activity obeys a homogeneous pattern, with the messages being slightly more evenly distributed in all lists than in simulations using uniform distribution[30]. In the networks, $\frac{max(incidence)}{min(incidence)} \in (1.26, 1.275]$ while simulations reach these values but have in average more discrepant higher and lower peaks $\xi = \frac{max(incidence')}{min(incidence')} \Rightarrow \mu_\xi = 1.2918$ and $\sigma_\xi = 0.04619$. Therefore, the incidence of messages at each second of a minute and at each minute of an hour was considered uniform, i.e. no trend was detected. Circular dispersion is maximized and the mean has little meaning as indicated in Table II. As for the hours of the day, an abrupt peak appeared around 11am with the most active period being the afternoon. Days of the week revealed a decrease of at least one third and at most two thirds of activity on weekends. Days of the month were regarded as homogeneous with an inconclusive slight tendency of the first week being more active. Months of the year revealed patterns matching usual work and academic calendars. The time period examined here was not sufficient for the analysis of ac-

TABLE III. Activity percentages along the hours of the day for the CPP list. Nearly identical distributions are found on other lists as shown in Section I C of the Supporting Information. Higher activity was observed between noon and 6pm, followed by the time period between 6pm and midnight. Around 2/3 of the whole activity takes place from noon to midnight. Nevertheless, the activity peak occurs around midday, with a slight skew toward one hour before noon.

	1h	2h	3h	4h	6h	12h
0h	3.66	6.42	8.20	9.30	10.67	33.76
1h	2.76					
2h	1.79	2.88				
3h	1.10		2.47			
4h	0.68	1.37		3.44		
5h	0.69					
6h	0.83	2.07	4.35		23.09	
7h	1.24					
8h	2.28	6.80		21.03		
9h	4.52		18.75			
10h	6.62	**14.23**				
11h	**7.61**					
12h	6.44	12.48	**18.95**	**25.05**	**37.63**	**66.24**
13h	6.04					
14h	6.47	12.57				
15h	6.10		18.68			
16h	6.22	12.58		23.60		
17h	6.36					
18h	6.01	11.02	15.88		28.61	
19h	5.02					
20h	4.85	9.23		17.59		
21h	4.38		12.73			
22h	4.06	8.36				
23h	4.30					

TABLE IV. Activity percentages along the days of the week for the four email lists. Higher activity was observed during weekdays, with a decrease of activity on weekends of at least one third and two thirds in extreme cases.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
LAU	15.71	15.81	15.88	16.43	15.14	10.13	10.91
LAD	14.92	17.75	17.01	15.41	14.21	10.40	10.31
MET	17.53	17.54	16.43	17.06	17.46	7.92	6.06
CPP	17.06	17.43	17.61	17.13	16.30	6.81	7.67

tivity along the years. These patterns are exemplified in Tables III-VI.

B. Scalable fat-tail structure: constancy of membership fractions in each Erdös sector

The distribution of vertices in the hubs, intermediary, periphery Erdös sectors is remarkably stable along time, provided that a sufficiently large sample of 200 or more messages is considered. Moreover, the same distribution applies to the networks of all email lists analyzed, as demonstrated in Figure 3 of current document and Section II of the Supporting Information. Activity is highly concentrated on the hubs, while a very large number of peripheral vertices contribute to only a fraction of the activity. This is expected for a system with a scale-free trace, as confirmed with the data in Table VII for the distribution of activity among participants.

Typically, $\approx [3-12]\%$ of the vertices are found to be hubs, $\approx [15-45]\%$ are intermediary and $\approx [44-81]\%$ are peripheral, which is consistent with the literature[31].

TABLE V. Activity in the days along the month for MET list. Nearly identical distributions are found on other lists as indicated in Section I E of the Supporting Information. Although slightly higher activity rates are found in the beginning of the month, the most important feature seems to be the homogeneity made explicit by the high circular dispersion in Table II.

	1 day	5	10	15 days
1	3.05			
2	3.38			
3	3.62	18.25		
4	4.25			
5	3.94			
6	3.73		35.24	
7	3.17			
8	3.26	16.98		50.96
9	3.56			
10	3.26			
11	3.81			
12	2.91			
13	3.30	15.73		
14	2.75			
15	2.95		31.98	
16	3.36			
17	3.16			
18	3.44	16.25		
19	3.36			
20	2.93			
21	3.20			
22	3.11			
23	3.60	15.79		49.04
24	2.74			
25	3.13		32.78	
26	3.13			
27	3.07			
28	3.61	16.99		
29	3.60			
30	3.57			

TABLE VI. Activity percentages of the months along the year from LAU list. Activity is concentrated in Jun-Aug for MET and LAD, and in Dec-Mar for CPP, LAU and LAD (see Section I F of the Supporting Information). These observations fit academic calendars, vacations and end-of-year holidays.

	m.	b.	t.	q.	s.
Jan	10.22	19.56	28.24	35.09	49.16
Fev	9.34				
Mar	8.67	15.53			
Apr	6.86		20.93		
Mai	7.28	14.07		30.36	
Jun	6.80				
Jul	8.97	16.29	24.47		50.84
Ago	7.32				
Set	8.18	16.25		34.55	
Out	8.06		26.36		
Nov	7.64	18.30			
Dez	10.66				

These results hold for the total, in and out degrees and strengths. Stable distributions can also be obtained for 100 or less messages if classification of the three sectors is performed with one of the compound criteria established in Section III C. The networks hold their basic structure with as few as 10-50 messages; concentration of activity and the abundance of low-activity participants take place even with very few messages, which is highlighted in Section II of the Supporting Information. A minimum window size for observation of more general properties might be inferred by monitoring the giant component and the degeneration of the Erdös sectors.

In order to verify the possible generality of these findings, we obtained the Erdös sectors of 12 networks from Facebook, Twitter and Participa.br. The results are given in Tables S30 in the Supporting Information, which indicate that the percentages of hubs, intermediary and peripheral nodes are essentially the same as for the email lists.

TABLE VII. Distribution of activity among participants. The first column presents the percentage of messages sent by the most active participant. The column for the first quartile ($1Q$) shows the minimum percentage of participants responsible for at least 25% of total messages. Similarly, the column for the first three quartiles $1-3Q$ gives the minimum percentage of participants responsible for 75% of total messages. The last decile $-10D$ column brings the maximum percentage of participants responsible for 10% of messages.

list	hub	$1Q$	$1-3Q$	$-10D$
LAU	2.78	1.19 (26.35%)	13.12 (75.17%)	67.32 (-10.02%)
LAD	4.00	1.03 (26.64%)	11.91 (75.18%)	71.14 (-10.03%)
MET	11.14	1.02 (34.07%)	8.54 (75.64%)	80.49 (-10.02%)
CPP	14.41	0.29 (33.24%)	4.18 (75.46%)	83.65 (-10.04%)

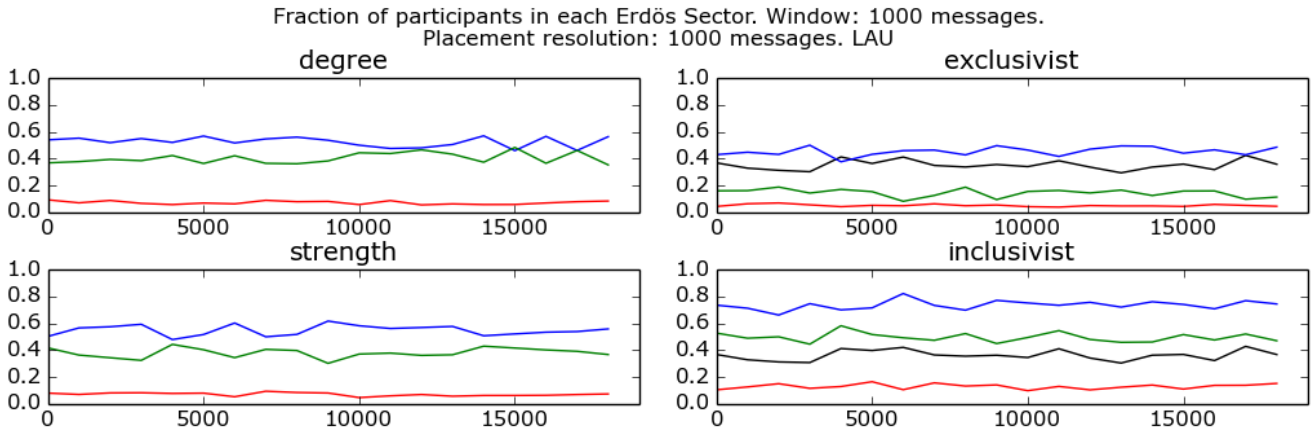


FIG. 3. Fractions of agents in each Erdös sector, where the fractions for hubs, intermediary and peripheral vertices are represented in red, green and blue, respectively. We used two simple criteria, namely degree and strength, for the graphics on the left. For the graphics on the right we employed the Exclusivist and Inclusivist compound criteria, with black lines representing the fraction of vertices without class and with more than one class, respectively. See Section II of Supporting Information for a collection of such timeline figures with all simple and compound criteria and metrics. Table S30, also from Supporting Information, presents these fractions of agents in snapshots of networks from Facebook, Twitter and Participa.br.

C. Stability of principal components and the prevalence of symmetry over clusterization for dispersion

The topology was analyzed using standard, well-established metrics of centrality and clustering. We also introduced symmetry metrics because of evidence of their importance in social contexts[7]. The contribution of each metric to the variance is very similar for all the networks, and did not vary with time. In applying PCA to the snapshots, the contribution of each metric to the principal components presents very small standard deviation. Table VIII exemplifies the principal components formation with all the metrics considered for the MET email list. Similar results are presented in Section III of the Supporting Information for the other lists, with separate consideration of strategic combinations of metrics.

The first principal component is an average of centrality metrics: degrees, strengths and betweenness centrality. Therefore, all of these centrality measurements are equally important for characterizing the networks. On one hand, the relevance of all centrality metrics is not surprising since they may be highly correlated. The degree and strength, for instance, are highly correlated, with Spearman correlation coefficient $\in [0.95, 1]$ and Pearson coefficient $\in [0.85, 1)$ for $ws > 1000$. On the other hand, each of these metrics is related to a different participation characteristic, and their equal relevance is noticeable. The clustering coefficient is presented in almost perfect orthogonality to centrality metrics.

Dispersion was more prevalent in symmetry-related metrics than for the clustering coefficient, as indicated in Table VIII. This is also illustrated in Figure 4, where each vertex is colored according to the sector they belong to. As expected, peripheral vertices have very low values in the first component (centrality related) and greater dispersion in the third component (clustering related). The PCA plot in the third system of Figure 4, where all metrics are considered, reflects the relevance of the symmetry-related metrics for the variance. We conclude that the latter metrics can be more meaningful in characterizing interaction networks (and their participants) than the clustering coefficient, especially for hubs and

TABLE VIII. Loadings for the 14 metrics into the principal components for the MET list, $ws = 1000$ messages in 20 disjoint positioning. The clustering coefficient (cc) appears as the first metric in the Table, followed by 7 centrality metrics and 6 symmetry-related metrics. Note that the centrality measurements, including degrees, strength and betweenness centrality, are the most important contributors for the first principal component, while the second component is dominated by symmetry metrics. The clustering coefficient is only relevant for the third principal component. The three components have in average 80% of the variance.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	0.89	0.59	1.93	1.33	21.22	2.97
s	11.71	0.57	2.97	0.82	2.45	0.72
s^{in}	11.68	0.58	2.37	0.91	3.08	0.78
s^{out}	11.49	0.61	3.63	0.79	1.61	0.88
k	11.93	0.54	2.58	0.70	0.52	0.44
k^{in}	11.93	0.52	1.19	0.88	1.41	0.71
k^{out}	11.57	0.61	4.34	0.70	0.98	0.66
bt	11.37	0.55	2.44	0.84	1.37	0.77
asy	3.14	0.98	18.52	1.97	2.46	1.69
μ_{asy}	3.32	0.99	18.23	2.01	2.80	1.82
σ_{asy}	4.91	0.59	2.44	1.47	26.84	3.06
dis	2.94	0.88	18.50	1.92	3.06	1.98
μ_{dis}	2.55	0.89	18.12	1.85	1.57	1.32
σ_{dis}	0.57	0.33	2.74	1.63	30.61	2.66
λ	49.56	1.16	27.14	0.54	13.25	0.95

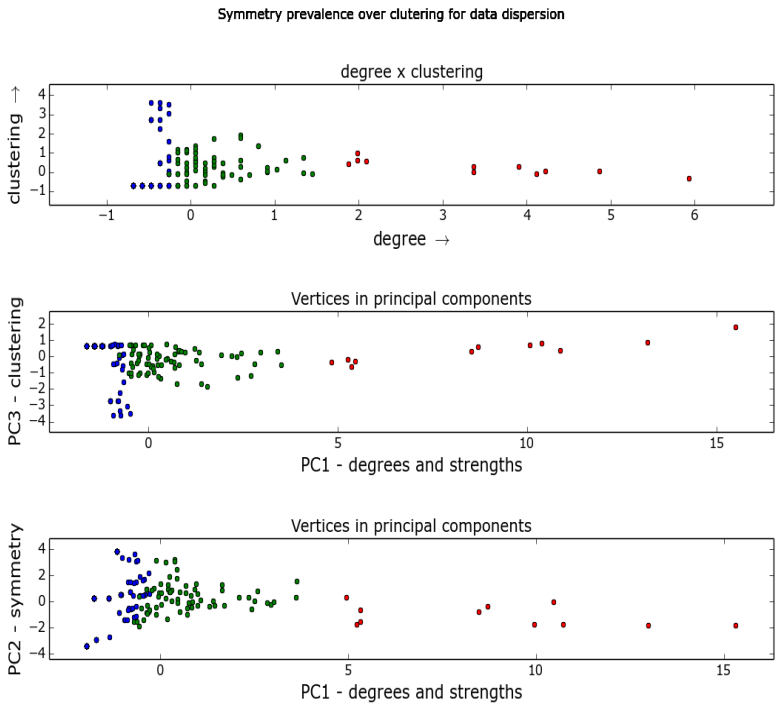


FIG. 4. The first plot shows degree versus clustering coefficient. This typical pattern is well known, since high clustering is more incident in vertices with lower degrees. The second plot is analogous but the first component is an average of centrality metrics. The second component remains related to the clustering coefficient. The third plot exhibits the greater dispersion in the symmetry-related second component. In this case, the clustering coefficient is only relevant for the third component. This greater dispersion suggests that symmetry-related metrics are more powerful for characterizing interaction networks than the clustering coefficient, especially for hubs and intermediary vertices. This figure was obtained with a snapshot of the LAU list in a window size of $ws = 1000$ messages. Similar structures were observed in all window sizes $ws \in [500, 10000]$ and for networks of other email lists, which points to a common relationship between the metrics of degrees, strengths and betweenness centrality, the symmetry-related metrics and clustering coefficient.

intermediary vertices.

The relative importance of the topological metrics was also observed for the additional 12 networks from Facebook, Twitter and Participa.br. With the exception of two of these networks, the overall behavior was maintained in that centrality measurements were found to be the most relevant to explain network topology, followed by symmetry-related metrics and then clustering coefficient. The results are given in Tables S31, S32, S33, S34 of the Supporting Information. There are larger differences between two of these networks than between two (GMANE) email networks, as the latter were much more regular.

D. Types from Erdös sectors

A sector to which a vertex belongs can be regarded as yielding a type to the corresponding participant. Assigning a type to a participant inevitably raises an important question regarding the possible stigmatization. We take the view that the participation typology inherent in the Erdös sectors is not stigmatizing because the type of an individual changes constantly[32]. That is to say, an individual is a hub in a number of networks and peripheral in other networks, and even within a network he/she probably changes type along time. Indeed, we did observe often transitions of participants from one sector to another. The typology proposed here bridges exact and human sciences and may be enriched with concepts from other typologies, such as Meyer-Briggs, Pavlov or the authoritarian types of the F-Scale[32].

We analyzed the time evolution of the networks using visualization tools developed for this research[33,34] and inspected the raw data to infer the main characteristics of each type. Our main observations may be summarized as follows:

- Core hubs usually have intermittent activity. Very stable activity was found on MET hubs, which is consistent with the literature where greater stability occurs in smaller communities[6].

- Typically, the activity of hubs is trivial: they interact as much as possible, in every occasion with

everyone. The activity of peripheral vertices also follows a simple pattern: they interact very rarely, in very few occasions. Therefore, intermediary vertices seem responsible for the network structure. Intermediary vertices may exhibit preferential communication to peripheral, intermediary, or hub vertices; can be marked by stable communication partners; can involve stable or intermittent patterns of activity.

- Some of the most active participants receive many responses with relative few messages sent, and rarely are top hubs. These seem as authorities and contrast with participants that respond much more than receive responses.

- The most obvious community structure, as observed by a high clustering coefficient, is found only in peripheral and intermediary sectors.

With regard to the networks as the whole objects of analysis, we were able to observe a negative correlation between the number of threads and the number of participants. When the number of participants exceeds a threshold, the number of threads displays a positive correlation with the number of participants. This finding is illustrated in Figure 5 and can also be observed in Table I. Obviously, network types can be derived from such results, which was not attempted here but left for the reader and future work.

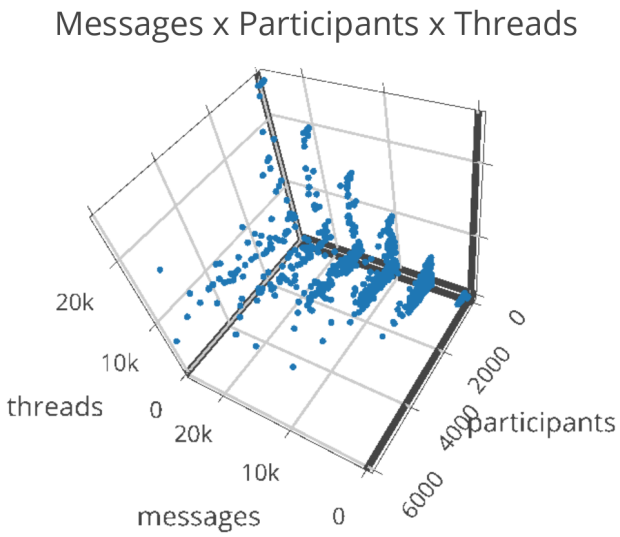


FIG. 5. A scatter plot of messages (M) versus participants (N) versus threads (Γ) for 140 email lists. Highest number of threads are found in lists with few participants. The correlation between N and Γ is negative for low values of N but positive otherwise. This negative correlation between N and Γ can also be observed in Table I. For $M = 20000$ messages, positive correlation of N and Γ is present mostly above 1500 participants. All LAU, LAD, MET lists present smaller networks.

V. CONCLUSIONS

The most important result from the analysis of time evolution of the four email lists is certainly the time-independence observed not only for the activity but also for the properties of the networks themselves. For example, the relative fractions of participants classified as hubs, intermediary and peripheral vertices remained practically constant along time, and this applied to all the email lists studied. Furthermore, the PCA analysis of the topological metrics characterizing the networks also indicated that the contribution of each metric did not vary in time. Centrality metrics were found to be the most relevant to characterize the network topology, followed by symmetry-related metrics, which were more relevant, with respect to variance, than clustering.

A systematic study of the activity of participants belonging to the three distinct Erdös sectors indicated simple patterns for hubs and peripheral vertices, while the network structure was governed by the intermediary vertices. These properties were shared by all email lists and were time-independent, being consistent with the literature. Moreover, both the distribution of Erdös sectors and the contribution from the metrics to the PCA were found to apply to networks from Facebook, Twitter and Participa.br. We may therefore consider the classification of agents into Erdös sectors as leading to a human typology which bridges between exact sciences, with objective procedures for the classification, and human sciences, where there is a legacy in the observation of human types.

ACKNOWLEDGMENTS

Financial support was obtained from CNPq (140860/2013-4, project 870336/1997-5), United Nations Development Program (contract: 2013/000566; project BRA/12/018) and FAPESP. The authors are grateful to the American Jewish Committee for maintaining an online copy of the Adorno book used on the epigraph[32], to GMANE creators and maintainers, and to the communities of the email lists and other groups used in the analysis, and to the Brazilian Presidency of the Republic for keeping Participa.br code and data open. We are also grateful to developers and users of Python scientific tools.

APPENDIX: DATA AND SCRIPTS

Messages were downloaded from the GMANE public database[16]. Data annotated from Facebook and Twitter are in a public repository[35]. Data from Participa.br was used from the linked data/semantic web RDF triples

reported in[36] and available in[37]. All routines necessary to achieve the results reported in this article, including all tables and figure of the Supporting Information, are available through a public domain Python package and an open Git repository[15]. Data from social networks used in this study were gathered and used within the Anthropological Physics principles[38].

[1] M. Newman, *Networks: an introduction* (Oxford University Press, 2010).

[2] B. Latour, "Reassembling the social. an introduction to actor-network-theory," Journal of Economic Sociology **14**, 73–87 (2013).

[3] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan, "Mining email social networks," in *Proceedings of the 2006 international workshop on Mining software repositories* (ACM, 2006) pp. 137–143.

[4] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási, "Modeling bursts and heavy tails in human dynamics," Physical Review E **73**, 036127 (2006).

[5] B. Ball and M. E. Newman, "Friendship networks and social status," arXiv preprint arXiv:1205.6822 (2012).

[6] G. Palla, A.-L. Barabási, and T. Vicsek, "Quantifying social group evolution," Nature **446**, 664–667 (2007).

[7] E. A. Leicht, G. Clarkson, K. Shedden, and M. E. Newman, "Large-scale structure of time evolving citation networks," The European Physical Journal B **59**, 75–83 (2007).

[8] B. Travençolo and L. d. F. Costa, "Accessibility in complex networks," Physics Letters A **373**, 89–95 (2008).

[9] M. E. Newman, "Modularity and community structure in networks," Proceedings of the National Academy of Sciences **103**, 8577–8582 (2006).

[10] K. Gergen and M. Gergen, *Historical social psychology* (Psychology Press, 2014).

[11] R. Albert and A.-L. Barabási, "Topology of evolving networks: local events and universality," Physical review letters **85**, 5234 (2000).

[12] K. Marek-Spartz, P. Chesley, and H. Sande, "Construction of the gmane corpus for examining the diffusion of lexical innovations," (2012).

[13] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis, "Geographic constraints on social network groups," PLoS one **6**, e16939 (2011).

[14] V. Palchykov, K. Kaski, J. Kertész, A.-L. Barabási, and R. I. Dunbar, "Sex differences in intimate relationships," Scientific reports **2** (2012).

[15] R. Fabbri, "Python package to observe time stability in the gmane database," (2015), https://pypi.python.org/pypi/gmane.

[16] Wikipedia, "Gmane — Wikipedia, the free encyclopedi," (2013), online; accessed 27-October-2013.

[17] Gmane.linux.audio.users is list ID in GMANE.

[18] Gmane.linux.audio.devel is list ID in GMANE.

[19] Gmane.comp.gcc.libstdc++.devel is list ID in GMANE.

[20] Gmane.politics.organizations.metareciclagem is list ID in GMANE.

[21] E. A. Leicht and M. E. Newman, "Community structure in directed networks," Physical review letters **100**, 118703 (2008).

[22] M. Newman, "Community detection and graph partitioning," arXiv preprint arXiv:1305.4974 (2013).

[23] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. Villas Boas, "Characterization of complex networks: A survey of measurements," Advances in Physics **56**, 167–242 (2007).

[24] M. O. Jackson, "Social and economic networks: Models and analysis," (2013), https://class.coursera.org/networksonline-001.

[25] I. Jolliffe, *Principal component analysis* (Wiley Online Library, 2005).

[26] U. Brandes, "A faster algorithm for betweenness centrality*," Journal of Mathematical Sociology **25**, 163–177 (2001).

[27] R. Fabbri, "Video visualizations of email interaction network evolution," (2013-5), https://www.youtube.com/playlist?list=PLf_EtaMqu3jVodaqDjN7yaSgsQx2Xna3d, https://www.youtube.com/playlist?list=PLf_EtaMqu3jWYQiJZYhVlJVngb7vsf6na, https://www.youtube.com/playlist?list=PLf_EtaMqu3jVb7CTt59t3ZnrmXuG0N3cO, https://www.youtube.com/playlist?list=PLf_EtaMqu3jVFS_AJZm_HuO9pywnSWaNF, https://www.youtube.com/playlist?list=PLf_EtaMqu3jU-1j4jiIUiyMqyVSzIYeh6, https://www.youtube.com/playlist?list=PLf_EtaMqu3jUZpAX3cKPC5J0t3q836CLy, https://www.youtube.com/playlist?list=PLf_EtaMqu3jUYO_XfJdqQELdbFnpqYEfb.

[28] R. Fabbri, "Image gallery of email interaction networks." (2013), http://hera.ethymos.com.br:1080/redes/python/autoRede/gmane.linux.audio.devel_3000-4200-280/.

[29] R. Fabbri, "Online gadget for making email interaction network images, gml files and measurements," (2013), http://hera.ethymos.com.br:1080/redes/python/autoRede/escolheRedes.php.

[30] Numpy version 1.6.1, "random.randint" function, was used for simulations, algorithms in https://pypi.python.org/pypi/gmane.

[31] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," Physics reports **424**, 175–308 (2006).

[32] T. W. Adorno, E. Frenkel-Brunswik, D. J. Levinson, and R. N. Sanford, "The authoritarian personality." (1950).

[33] R. Fabbri, "A connective differentiation of textual production in interaction networks," (2013), http://arxiv.org/abs/1412.7309.

[34] R. Fabbri, "Versinus: a visualization method for graphs in evolution," arXiv preprint arXiv:1412.7311 (2013), http://arxiv.org/abs/1412.7311.

[35] R. Fabbri, "Python package to analyze the gmane database," (2015), https://pypi.python.org/pypi/gmane.

[36] R. Fabbri, "Content extraction through api from the Brazilian Federal Portal of Social Participation and its tools to a social participation cloud," Tech. Rep. (United Nations Development Programme and Brazilian Presidency of the Republic, 2014) https://github.com/ttm/pnud5/blob/master/latex/produto.pdf?raw=true.

[37] R. Fabbri, "Data from Participa.br, Cidade Democrtica and AA, in XML/RDF and Turtle/RDF," (2014), http://datahub.io/organization/socialparticipation.

[38] R. Fabbri, "What are you and i? [anthropological physics fundamentals]," (2015), https://www.academia.edu/10356773/What_are_you_and_I_anthropological_physics_fundamentals_.