Analítica Visual (Visual Analytics) para processos de difusao em redes complexas

Proponente: Dr. Renato Fabbri
Supervisora: Profa. Dra. Maria Cristina Ferreira de Oliveira
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo (ICMC-USP)
24 de maio de 2020

Resumo: Os processos de difusao sao modelados de diversas formas, e o estudo destes processos em redes eh uma area ampla. Englobam e.g. modelos inspirados em e aplicacoes para epidemiologia, redes de relacoes entre genes e proteinas, de informacao em estruturas sociais humanas, de estado fisico ou de sistemas computacionais. Este projeto propõe o desenvolvimento de métodos e software de analítica visual (*visual analytics*) para processos de difusao em redes complexas. Em especial, ha propostas de modelos potencialmente ineditos que podem ser estudados com o auxilio de simulacoes visuais, mesmo nos casos em que os metodos numericos e a analitica algebrica e torna-se custosos, dificeis ou de exposicao laboriosa. O proponente deste plano de trabalho colabora com o VICG/ICMC/USP, desde que iniciou o pos-doutorado adiante adaptado para um Treinamento Tecnico V (FAPESP XXXX), e pesquisou a caracterizacao de redes complexas sociais no seu doutorado. Assim, ha artigos e ferramentas desenvolvidas atraves dos quais o pesquisador retomarah a pesquisa e estabelecerah coesao dentre as pesquisas realizadas e propostas. Por exemplo, foram concebidos metodos e ferramntas de visualizacao de redes bipartidas e de redes logitudinais (i.e. evolutivas, dinamicas; que ganham e perdem vertices e arestas em uma sucessao de eventos). Destacam-se os metodos e ferramentas para visualizacao de redes bipartidas assistidos por estrategias multinivel. Ao incrementar as contribuicoes jah efetivadas, algumas interfaces e metodos em software, e artigos, estarao em melhor estado para continuidade e colaboracao em outras parcerias e orientacoes academicas.

Palavras-chave: Analitica visual, Redes sociais, Redes complexas, Mineração de texto.

Visual Analytics of text and topology in social networks

Proponent: Dr. Renato Fabbri
Supervisor: Prof. Dr. Maria Cristina Ferreira de Oliveira
Institute of Mathematics and Computational Sciences,
University of São Paulo (ICMC-USP)

24 de maio de 2020

Abstract:
Keywords: Visual analytics

1 Formulacao do problema

"Essentially, all models are wrong, but some are useful"(George Box, 1987)

1.1 Aspectos eticos

Nos trabalhos [] estao consideradas questoes eticas da pesquisa cientifica envolvendo seres humanos. Em especial, foram desenvolvidos emprestimos da antropologia para a pesquisa em fisica, sob o acompanhamento de academicos mais experientes, principalmente e nominalmente: Profa. Dra. Marilia Pisani (Filosofia, UFABC), Profa. Dra. Deborah Antunes (Psicologia, UFC), e Prof. Dr. Massimo Canevacci (Sapienza Universita di Roma). Em resumo, os aspectos eticos foram amenizados atraves do estudo das estruturas sociais do proprio pesquisador (emprestimo da "escrita de diarios", tecnica da antropologia etnografica), e da manutencao da transparencia da pesquisa com manutencao para acesso publico de textos, dados e software (emprestimo da cultura livre: contribuicao para o legado publico da humanidade de conhecimentos e tecnologias). Com isso, foi possivel realizar alguns experimentos nas redes do proprio pesquisador, e a consideracao destes experimentos em documentos cientificos, sem o investimento de tempo para vencer burocracias e procedimentos de comites de etica.

No ambito deste plano de trabalho, o pesquisador concentrara esforcos no desenvolvimento de modelos, ferramntas em software, simulacoes, e relatoria cientifica. Este aspecto da pesquisa nao implica na necessidade da aprovacao de qualquer comite de etica. Mesmo assim, caso haja espaco e pertinencia para experimentos em estruturas sociais, as propostas dos experimentos serao consideradas pelos pesquisadores proponente e pela responsavel para submissao aos comites de etica cabiveis.

1.2 Sincronizacao

Simplificadamente , a sincronizacao eh tratada em conjunto com ou como um efeito permanente na rede [**?**]. Um pouco por reuso do vocabulario e pelos modelos de interesse, chamamos aqui de sincronizacao uma classe de processos e modelos de difusao.

Alguns modelos desta classe foram criados para desenvolvimento tecnologico pessoal do pesquisador, e visa a propagacao de informacao e concordancias sociais. Considere um grafo (i.e. uma rede) $G = (V, E)$, em que $V = \{v_i\}_0^{N-1}$ sao vertices, e $E = \{e_i = (v_j, v_k)\}_{i=0}^{z-1}$ sao arestas. Os pressupostos gerais sao:

- alguns vertices v_i, $|v_i| = a << N$, sao ativados para iniciar a propagacao, chamados sementes ou vertices iniciais.

- Cada vertice, ao ser ativado, ativa b (constante) vizinhos e nao pode mais ser ativado. Caso o vertice nao possua b vizinhos nao-ativados, ativara todos os vizinhos possivel.

Os criterios de escolha das sementes, a, b, sao arbitrarios. Alem disso, ha variantes, e.g. os vertices ativados podem ser novamente ativados, ou b pode ser variante no tempo, randomico ou dependente de caracteristicas dos vertices (e.g. grau). Pode-se tambem impor restricoes adicionais, dentre os quais a mais comum eh que todos os vertices sejam ativados (sincronizacao completa).

As experiencias descritas em [**?**, **?**, **?**, **?**, **?**] resultaram na imposicao de pouca atividade de cada vertice, e propagacao desta atividade, pois demoraram meses alguns procedimentos (e.g. os realizados em Dez/2012 ateh Mar/2013). Os experimentos mais efemeros eram baseados em comunicacao (i.e. ativacao) de alguns poucos vertices e nao trouxeram mudanca nitida na estrutura e funcionamento da rede (assuntos, trocas, etc). Tambem foi obtida a heuristica de comecar pelos vertices menos conectados, os perifericos [**?**, **?**]. O detalhamento das justificativas e evidencias que corroboram estas caracteristicas foge ao escopo deste documento e encontra-se na literatura citada.

Por fim, pode-se utilizar tecnicas avancadas de ciencia de redes. Considere o modelo basico resultante dos pressupostos gerais e consideracoes acima. Ou seja, a sementes, b ativacoes por ativacao,

Para exemplificar um processo de sincronizacao, considere uma estrategia multinivel, em uma rede original G_0 eh representada como uma sucessao de redes $M = \{G_i\}_0^{m-1}$ tal que se $i < j$, $N_i \leq N_j$ e $z_i \leq z_j$. Para a obtencao de G_j a partir de G_i, sao determinados os conjuntos de vertices que serao colapsados em supervertices atraves de algum dentre os varios algoritmos de *matching*. V_j resulta dos supervertices e dos vertices que nao fizeram parte de nenhum conjunto colapsado. E_j resulta das arestas implicadas pela rede G_i e pela correspondencia entre V_i e V_j, um algoritmo que depende das caracteristicas da rede (simples, com peso, direcionada, bipartida, multicamada, heterogenea, etc).

Com o objetivo de obter sementes iniciais e a sequencia de vizinhos a serem ativados,

Um conjunto de vertices para colapso consistirah na escolha dos b vizinhos mais conectados de um vertice. Este vertice serah o vertice mais conectado que ainda nao pertence a um destes conjuntos. O processo continuarah ate na rede menor restem somente b vertices que nao participaram de nenhum colapso.

A introducao de ruido na escolha dos vertices e vizinhos a serem colapsados

permite a comparacao e escolha entre as diferentes colecoes de sementes $S = \{v_i\}_0^{s-1}$ e arestas A relacionadas a cada vertice $A = \{(v_i, \{v_j\})_0^{a_i-1}\}_0^{a-1}$ a serem exercitadas na propagacao da ativacao.

Na propagacao de um patogeno, o contagio eh muitas vezes modelado sendo igualmente possivel por cada aresta. Este eh realmente o caso quando o patogeno eh assintomatico. Ja em quadros mais complicados e obitos, o individuo pode contagiar diretamente apenas um numero limitado de pessoas, pois retira-se para tratamento ou internacao. Alem disso, dado que a rede observada eh integrada a outras redes, o patogeno eh introduzido na rede G por redes nao observadas, sem a necessidade de arestas em E.

Estas ultimas consideracoes sao uteis para estudos epidemiologicos (e.g. SARs-Covid-19). O modelo em que A eh determinado adequa-se melhor para aplicaoes comerciais (e.g. Marketing Multinivel) e para *crowdsourcing* de infomacao/dados (e.g. democracia liquida e formacao de redes).

1.3 Trabalhos relacionados de outros autores

Os processos de difusao em redes complexas tem recebido crescente atencao na literatura cientifica. De fato, os modelos sao utilizados para uma classe vasta de fenomenos que ganharam mais relevancia na ciencia recente, como redes de interacoes entre proteinas e genes em celulas, de contagio na epidemiologia, de informacao, opiniao e fofoca em plataformas e estruturas sociais.

Eh possivel, neste contexto, discerninr duas abordagens paradigmaticas. Ha a abordagem abstrata, em que os vertices e arestas nao sao definidos para alem de suas definicoes matematicas ou em que sao minimizadas as particularidades dos sistemas de interesse. Ha abordagens bastante especializadas para os sistemas de interesse (e.g. biologico, social, fisico, tecnologico), e nesta linha provavelmente se destacam as "analises de difusao baseadas em redes" e as redes medicas.

Interessa-nos em especial adaptar os modelos para utilizacao de estrategias multinivel, e redes em que ha uma modificacao sequencial nos vertices e arestas considerados. A caracterizacao de novos modelos, como o descrito na secao anterior, seguirah a pertinencia como reforcada pelos trabalhos dos pesquisadores envolvidos ou por lacunas encontradas na literatura.

Como esperado, o propontente acompanha a pesquisa em andamento na ciencia da complexidade, redes complexas, e redes de difusao. Este acompanhamento inclui os artigos e livros da bibliografia, e fontes que se mostraram uteis no decorrer da pos-graduacao e projeto FAPESP (). Tais fontes incluem MOOCs[1],

[1] a

software[2], assinatura e contribuicoes em paginas da Wikipedia, e notas publicadas por instituicoes competentes[3]

2 Objetivo

O objetivo deste plano de trabalho eh o desenvolvimento de modelos de processos de difusao em redes, e a implementacao de interfaces de visualizacao

2.1 Objetivos especificos

3 Justificativa

3.1 Historico de pesquisa especializada do proponente

4 Metodologia

O trabalho sera acompanhado de constante aprofundamento teorico na area. Com base na experiencia do pesquisador na area, ha pertinencia na visita a MOOCs[4], e.g. *Network Dynamics of Social Behavior* (https://www.coursera.org/learn/networkdynamics/home/welcome), e este segmento do *Curso do cara da stanford* (https://saoid).

Serao garantidos um minimo de 12 horas de dedicacao semanais com foco absoluto neste plano de trabalho, a serem distribuidos entre aprofundamentos teoricos, desenvolvimento de metodos e modelos, desenvolvimento e manutencao de software, articulacao com outros pesquisadores e instituicoes, e escrita de artigos e relatorios cientificos. Este eh o numero minimo de horas semanais previstas na Resolução CoPq Nº 7413. O pesquisador estarah mantendo outras atividades de pesquisa e engenharia de software relacionadas aa ciencia de redes, habito mantido ha mais de 10 anos. Caso novas tecnologias e metodos desenvolvidos sejam diretamente relacionados a este plano de trabalho, o pesquisador entrarah em contato com a Pro-Reitoria de Inovacao, como previsto na resolucao supracitada.

5 Cronograma de execucao

1. Concepcao do plano de trabalho e estabelecimento da parceria.

[2] graphology, networkx
[3] https://www.santafe.edu/research/projects/transmission-sfi-insights-covid-19.
[4] Sigla de *Massive Online Open Courses*

2. Aprofundamento do conhecimento de teoria de redes complexas e processos de difusao em redes, em conformidade com a Secao .

3. Acréscimos aos modelos atuais de analítica visual e visualização de dados aplicados à análise de redes sociais, com o foco no participante da rede, nos pesquisadores em potencial e na classificação/tipologia de redes e participantes.

4. Implementação computacional. Estamos já implementando layouts para grafos no ccNetViz.

5. Escrita e publicação dos resultados em artigos. Esta etapa está já em andamento pois possuímos diversos escritos com resultados relacionados à mineração e visualização de dados de redes sociais que estão sendo submetidos para publicação.

6. Trocas com pesquisadores externos, estabelecimento de colaborações.

7. Elaboração do relatório científico final.

	2017	2018		2019
Atividade	2°	1°	2°	1°
1	•	•		
2	•	•		
3	•	•	•	•
4	•	•	•	•
5	•	•	•	•
6				•

Tabela 1: Cronograma de atividades ao longo dos semestres.