Analítica Visual (Visual Analytics) de Topologia e Texto em Redes Sociais

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

22 de maio de 2017

Resumo:
Palavras-chave: Visual analytics, Redes sociais, Redes complexas, Mineração de texto.

Analítica Visual (Visual Analytics) de Topologia e Texto em Redes Sociais

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

22 de maio de 2017

Abstract
Keywords: Visual analytics, Social networks, Complex networks, Text mining.

1 Enunciado do problema

Este trabalho se propõe a desenvolver métodos e ferramentas para a Análise de Redes Sociais (ARS) via Analítica Visual (*visual analytics*), orientados à mineração destas estruturas por especialistas de outras áreas (e.g. linguístas, físicos, cientistas sociais, psicólogos) e considerando as propriedades das estruturas sociais. Exemplos destas propriedades podem ser divididos em três grupos:

- Topológicos: as redes sociais são reportadas como livres de escala, de pequeno mundo e com comunidades [**?**].
- Textuais: a linguagem nas redes sociais é muitas vezes informal, com abreviações e erros com relação às normas oficiais [**?**].
- Relacionamento entre topologia e texto: há traços do texto produzido pelos participantes que se relacionam à topologia. Por exemplo, hubs utilizam frases mais curtas e mais verbos e advérbios, periféricos utilizam mais substantivos [**?**].

Em resumo, nossa proposta de desenvolver uma ferramenta de analítica visual para redes sociais envolve a integração dos dados a serem analisados, o processamento dos dados para obtenção de estruturas relevantes, da concepção de visualizações e formas de interação com elas, e da escrita de software que permita aos pesquisadores de outras áreas usufruir destas rotinas através de uma interface unificada.

1.1 Terminologia, polissemia e sinônimos

A pesquisa proposta neste documento envolve várias áreas do conhecimento. O conhecimento pode ser sistematizado de diversas formas. Entendemos que esta divisão é útil para o trabalho:

- Analítica Visual (*visual analytics*): compreendida aqui como a mineração de dados através de visualização de dados em interfaces interativas.
- Análise de Redes Sociais (ARS): realizada a partir das áreas de redes complexas e mineração de texto.

Salta evidente que as duas áreas são também multidisciplinares. As áreas diretamente relacionadas pela nossa descrição são também multidisciplinares: mineração de dados, visualização de dados, interfaces interativas, redes complexas, mineração de texto. Além disso, as áreas são recentes. As redes complexas surgiram como área de investigação reconhecida no final da década de 90 [**?**]; os primeiros workshops de mineração de texto ocorreram em 1999 [**?**]; como área independente, o desenho de grafos (*graph drawing*) surgiu na década de 1990 [**?**]. Ainda não citado, mas importante para a forma como abordamos alguns dos desafios da pesquisa proposta, o termo "dados ligados" foi cunhado em 2006 [**?**]. Como resultado, o vocabulário não é sempre bem definido, há palavras diversas

que se referem aos mesmos conceitos e conceitos diferentes relacionados a uma mesma palavra.

Uma apresentação detalhada da terminologia, incluindo escolhas consistentes de terminologia e os diferentes problemas nas terminologias encontradas, foge ao escopo deste projeto de pesquisa. Para exemplificar, na área de redes complexas, os termos "rede" e "grafo" são utilizados muitas vezes de forma indistinta, embora em textos mais cuidadosos a palavra grafo se refira à estrutura matemática composta de vértices e arestas (relacionamentos binários/duais entre vértices) e rede em geral se refira a grafos encontrados em sistemas reais/empíricos/naturais ou ao sistema representado como grafo. Outro bom exemplo é a usual indistinção entre mineração de texto, linguística computacional e processamento de linguagem natural (PLN). Uma distinção coerente entre as áreas pode ser feita considerando a mineração de texto como a mineração de dados aplicado a textos, a linguística computacional como a modelagem (estatística ou orientada a regras) de linguagem natural, e o PLN como a pesquisa e desenvolvimento aplicados à interação entre computadores e a linguaagem humana.

Outros termos e áreas importantes para o trabalho proposto são: interação humano-computador, percepção, cognição, estímulo, representação, representação de dados (de uma, duas, três ou mais dimensões), visualização (de dados, de informação, científica), raciocínio analítico (*analytical reasoning*), reconhecimento de padrões, aprendizado de máquina, design, design de interação, design de interfaces, cores, formas, glifos, analítica visual (*visual analytics*), análise de tipos específicos de dados (áudio, imagem, estruturas sociais, dados sintéticos).

Talvez a conceitualização mais útil nesta pesquisa, que transcende o propósito de expor inconsistências ou a busca vocabulários e definições concisos, é o estudo da classificação. Nas humanas, há tradição no estudo de *tipologias*, que são classificações em geral não bem definidas quantitativamente e mais preocupadas em qualificar os diferentes tipos ou classes. Já as *taxonomias* são em geral reconhecidas como de grande tradição na biologia, e possuem ênfase na nomenclatura com base em características bem definidas. A definição de classificação, taxonomia e tipologia não é uniforme na literatura científica (e.g. árvore taxonômica na linguística se refere ao relacionamento dos hiperônimos aos hipônimos e não se referem à biologia) e há outros termos referentes à classificações: vocabulários, ontologias, tesaurus, etc. Ainda assim, a consideração da literatura sobre classificação em diferentes áreas do conhecimento poderá ajudar-nos a uma melhor organização do conhecimento e a contemplar as necessidades de pesquisadores com formações diversas.

1.2 Sua importância

Há uma proliferação de redes sociais virtuais que produzem muitos dados e as estruturas relacionadas estão ainda sendo caracterizadas. Uma ferramenta de analítica visual em cógido aberto, portável para sistemas operacionais diferentes, e que seja capaz de lidar com grandes quantidades de dados é certamente algo desejável. As análises a serem possibilitadas pela ferramenta, e necessárias para a concepção da mesma, são úteis para aprofundar nossa compreensão das redes

sociais, principalmente virtuais.

Na pesquisa de doutoramento do candidato deste projeto de pesquisa, foram feitas contribuições relevantes à caracterização destas estruturas sociais a partir das áreas de redes complexas e de mineração de texto. O trabalho aqui proposto visa permitir que estas caracterizações sejam aplicadas a um conjunto maior de redes virtuais para melhor delimitar sua validade. Ao mesmo tempo, este conhecimento sobre as estruturas analisadas possuem um potencial para classificações e tipologias nas redes e nos participantes. Além da relevância evidente das redes sociais virtuais, há interesse neste estudo de tipos (de redes e de participantes) de nossa parte, da parte da Rede Nexos de Pesquisa Interdisciplinar [**?**] (do qual o proponente faz parte há anos) e da tradição de exatas em geral, que tem grande ênfase na tarefa de classificação.

Melhor relacionando ARS com redes complexas e mineração de texto, podemos esperar que haja contribuições na análise de literatura e de redes livres de escala. Este aspecto não é central para a proposta mas mesmo assim ainda relevante pois, para melhor permitir o estudo das redes sociais, é útil que haja outras redes e textos para comparação, o que permitirá a ARS em relação a estas estruturas externas e vice-versa.

1.3 A contribuição se bem sucedido

Do ponto de vista dos métodos, as análises pertinentes, assim como escolhas de interfaces apropriadas, já resultaram em publicação científica pelo proponente [**?**] e pela supervisora [**?**, **?**] e deverá implicar em maiores contribuições para a compreensão da analítica visual e das estruturas sociais virtuais. O software a ser desenvolvido certamente é uma contribuição pois possibilitará avanços na analítica visual e na caracterização dos sistemas sociais e potencialmente também de literatura e redes livres de escala.

1.4 Trabalhos relevantes

A tese de doutoramento do proponente *"Topologial stability and textual differentiation in human interaction networks: statistical analysis, visualization and linked data"* apresenta contribuições para a caracterização dos sistemas sociais virtuais, algumas destas em processo de publicação [**?**, **?**, **?**] e uma já aceita pela revista Physica A [**?**]. Há trabalhos do grupo de pesquisa em visualização de dados e de redes [**?**, **?**, **?**]. Os trabalhos [**?**, **?**, **?**] são base para visualização de dados, enquanto [**?**, **?**, **?**] são especificamente para visualização de redes/grafos e [**?**, **?**, **?**] são sobre visualização para mineração de texto. Os trabalhos [**?**, **?**, **?**] são bastante relevantes sobre visual analytics e [**?**, **?**] são referência sobre interação humano-computador. Há diversos software que deverão ser utilizados no decorrer do trabalho para apreender as possibilidades de analítica visual neste contexto (e.g. Gephi, Weka, ccNetViz, D3.js, Three.js). Há também literatura e artefatos mais ligados à arte, como a teoria da Gestalt, kiki-bouba, design e estética [**?**, **?**, **?**, **?**].

1.4.1 Projeto de Pesquisa Regular relacionado a esta proposta de pós-doutoramento

A supervisora desta proposta, a Profa. Dra. Maria Cristina Ferreira de Oliveira, está pesquisando e possui um aluno de mestrado no tema. Ela enviou recentemente uma proposta de projeto temático para a FAPESP. Segue o resumo.

"A pesquisa em Visual Analytics é central no tratamento dos desafios associados à análise de dados e computação intensiva em dados, pelo potencial de combinar técnicas de Aprendizado de Máquina e de Visualização para apoiar a interpretação de dados complexos. O acoplamento de técnicas oriundas de ambas as áreas pode promover avanços significativos na capacidade humana de análise de dados, pois permite a indivíduo e computador assumirem papeis complementares ao tratar os muitos problemas introduzidos pelo volume e complexidade dos conjuntos de dados gerados em diversos domínios de aplicação. Este projeto de pesquisa aborda dois focos distintos em visual analytics, um de natureza aplicada e outro de natureza conceitual. No aspecto aplicado serão considerados (i) o problema de visualização de redes de grande escala, com ênfase em redes sociais; e (ii) o problema da análise exploratória de espaços de atributos que caracterizam fenômenos multivariados e variantes no tempo - por exemplo, resultantes de sensores utilizados para monitoramento ambiental em diversos domínios. Em ambos os casos, a busca por soluções escaláveis para grandes volumes de dados representa um desafio. No aspecto conceitual, dando continuidade a uma colaboração em andamento, iremos conceber e realizar alguns estudos experimentais que contribuam para esclarecer os processos cognitivos subjacentes à interpretação de um tipo particular de visualização multidimensional, os chamados mapas de similaridade. A análise dos resultados pode sugerir modelos conceituais sobre a interpretação desse tipo de mapeamento visual. Esperamos com esse estudo contribuir para ampliar o embasamento conceitual sobre essas técnicas, essencial para futuros avanços na área."

Ou seja, há um interesse do grupo VICG (*Visualization, Imaging and Computer Graphics*) na ARS através da analítica visual que é anterior à concepção desta proposta de pós-doutoramento. De forma bastante conveniente, o trabalho de doutoramento do candidato fornece um arcabouço de análise, dados já representados como dados ligados e visualizações iniciais, o que permite que esta proposta de pós-doutoramento seja realizada com maior facilidade.

2 Resultados esperados

São esperados resultados em três frentes:

- Melhor caracterização das redes sociais virtuais através de mineração da topologia e do texto.
- Avanços em analítica visual de redes sociais principalmente através da consideração das características topológicas e textuais para a interface.

- Um software aberto para analítica visual de redes sociais que seja capaz de lidar com grandes volumes de dados e seja executado em navegadores web.
- Um arcabouço de analítica visual que permita comparar redes sociais com outras redes reais (especialmente livres de escala) e com textos diversos (e.g. literatura científica, James Joyce, Shakespeare, Bíblia).
- Um arcabouço de analítica visual que facilite estudos classificatórios (tipológicos, taxonômicos) em redes sociais de forma a permitir a observação de tipos/classes de redes e de participantes por pesquisadores e.g. de ciências e psicologia sociais.

2.1 Disseminação dos resultados

O software aberto pretendido deve facilitar a interação com usuários, desenvolvedores e pesquisadores. Visamos que o software seja utilizado por pesquisadores de áreas diversas como linguística, física e ciências sociais. Os resultados devem ser apresentados em conferências e em artigos de revistas, principalmente internacionais. A circulação de material didático em video potencialmente aumentará a recepção das contribuições. As análises e interfaces devem ser devolvidas às comunidades virtuais que geraram os dados.

3 Desafios científicos e tecnológicos

A consideração da tarefa revela alguns desafios. Primeiro, a análise de texto e de topologia implicam em um conjunto de medidas de alta dimensionalidade. Segundo, o montante de dados a serem analisados é bastante grande e as medições são muitas vezes computacionalmente caras. Por exemplo, a medida usual de *betweenness centrality* requer que sejam encontradas todas as geodésicas entre todos os pares de vértices [**?**]; as medidas de texto são resultantes de muitos dados e podem envolver procedimentos não triviais como a etiquetação morfossintática ou similaridades via rede semântica (Wordnet) [**?**].

Este projeto propõe a utilização destas análises, e aprofundamento delas, para a analítica visual. Os desafios imediatos para esta tarefa são:

- O uso apropriado de técnicas de visualização. Os layouts para redes orientados a força são os mais usuais em análise de redes sociais por explicitarem comunidades e resultarem em figuras esteticamente atraentes, mas a observação de outras características ficam prejudicadas, um fenômeno que tem recebido o nome de *hairball effect*. Para a visualização de dados multidimensionais, as técnicas podem não ser apropriadas para dados com dimensionalidade muito alta ou não resultarem em coordenadas com fácil interpretação, como no caso do LSP. Além disso, o uso de cores, formas e glifos possuem um alto impacto na capacidade de análise pelo usuário e

possuem muitas teorizações envolvidas e são substancialmente modificadas de acordo e.g. com o monitor do usuário, o que dificulta o controle na implementação.

- Viabilizar interatividade apropriada com o montante de dados e medições exposto acima. Por exemplo, quais controles devem ser disponibilizados e com quais controladores? Como compatibilizar a interatividade e a quantidade de dados a serem analisados?
- Possibilitar a análise de estruturas arbitrárias, i.e. de redes formadas por quantidades arbitrárias de mensagens de emails ou da junção de diferentes redes do Facebook.
- A escrita de software multiplataforma que lide com a integração dos dados, processamento destes para obtenção de estruturas relevantes, visualização e interatividade.

3.1 Meios e métodos para superá-los

Desafios da análise de topologia e texto em redes sociais foram abordados pelo proponente em [**?**], resultando em caracterização das estruturas sociais quanto à estabilidade topológica e diferenciação textual. Para os outros desafios ligados explicitamente à analítica visual, itemizados acima, propomos:

- A utilização de layouts geometricamente inspirados, como os *Hive Plots* [**?**] ou o Versinus (este desenvolvido pelo proponente [**?**]). A possibilidade de utilização de layouts e plots diferentes e de layouts simulâneos em regiões distintas da interface. Há também a previsão do uso de técnicas de redução de dimensionalidade (e.g. MDS, LSP) dada a grande quantidade de medidas relevantes para analisar redes sociais quanto à topologia e ao texto. Outras visualizações já bem estabelecidas devem ser disponibilizadas e.g. plot quantil-quantil (Q-Q) e de cumulativas de distribuições para viabilizar a análise dos dados de natureza estatística.
- O estudo da literatura da área nos fornece pistas valiosas, mas certamente só saberemos quais as reais possibilidades de interatividade através da implementação e de testes subsequentes. Prevemos também que, evidentemente, as possibilidades de interatividade de processamento massivo de dados dependerá do hardware do usuário, uma vez que deverá ser executado no cliente ao menos a visualização.
- Para que as estruturas a serem analisadas possam ser integradas e observadas por partes, podemos aproveitar a representação de estruturas sociais como dados ligados já desenvolvida pelo proponente [**?**, **?**]. A base de dados resultante (já obtido) preferencialmente deve ser abrigado em alguma instancia publicamente acessível e pode ser expandida no decorrer da pesquisa proposta neste projeto. Uma possibilidade gratuíta é utilizar os serviços do Data-World [**?**] pois abrigam dados ligados, mesmo na grande quantidade que temos, e permitem consultas via SparQL.

- Entendemos que o ideal é que o software possa ser executado em navegadores web usuais (e.g. Google Chrome e Firefox). Para isso, contamos com boas bibliotecas de visualização. A principal é a D3.js, que pode ser integrada com WebGL para possiblitar visualizações mais complexas. Boas alternativas são o Three.js e o ccNetViz, que já são integrados com WebGL. Já no processamento dos dados, podemos aproveitar o bom repertório de bibliotecas em Python, que utiliza as rotinas em BLAS e LAPACK para processamento matricial e vetorial. Devemos estudar as possiblidades de implementar o processamento dos dados diretamente em JavaScript para que a arquitetura do software fique simplificada, mas isso não é central, dado que a integração de JavaScript com Python é imediata caso sejam utilizados frameworks em Python para desenvolvimento web como Flask e Django. A maior diferença neste caso é que a implementação das rotinas de processamento dos dados em Python serão executadas pelo servidor ao passo que, se implementadas em JavaScript, poderão ser executadas no cliente (caso seja feita uma implementação em Node.js, o processamento em JavaScript poderá ser feito no servidor). Concebemos toda a implementação em código aberto/software livre para facilitar a colaboração no desenvolvimento e permitir que as rotinas sejam publicamente escrutinizadas.

3.2 Adequação do candidato à proposta

O candidato a esta pesquisa possui uma formação altamente multidisciplinar, o que deverá contribuir para o bom andamento do trabalho. Expandindo o argumento:

- a pesquisa envolve muitas áreas de conhecimento, de questões ligadas ao audiovisual e percepção, a análise de texto e redes e engenharia de software, como exposto na Seção 1.1.
- O candidato desenvolveu trabalhos nas áreas envolvidas: é graduado em artes (composição musical); realizou mestrado e doutorado na física computacional com contribuições científicas de nítido cruzamento entre física, estatísica, mineração de texto, redes complexas, ontologias e dados ligados, programação, e artes [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. Além disso, realizou diversas atividades de cunho prático, como composição musical, apresentações artíticas, programação de software original e contribuição para diversos software livres bastante utilizados no mundo todo, e uma consultoria em parceria com a ONU e a Presidência da República [**?**, **?**, **?**].

Em resumo, o candidato está acostumado a lidar com as diferentes linguagens e a grande quantidade de literatura envolvida em um trabalho como este. Adicionalmente, possui algum conhecimento e maturidade com relação às áreas, como evidenciado pelo método de visualização de redes em evolução que desenvolveu [**?**], pela mineração de dados realizada no seu doutorado [**?**], a modelagem psicofísica apresentada em seu mestrado [**?**, **?**], e as implementações computacional que permeiam praticamente toda a sua produção.

4 Cronograma

24 meses.

5 Disseminação e avaliação

Como exposto na Seção 2.1, haverá disseminação das contribuições em conferências, revistas, video e através do software em código aberto. O desenvolvimento deverá estar publicamente acessível desde o começo, muito provavelmente em repositório Git na plataforma Github. A avaliação deverá ser feita em três frentes: pela nossa própria capacidade de mineração dos dados, idealmente para obtenção de resultados suficientemente relevantes para publicação; através de testes com usuários em potencial (e.g. da Rede Nexos de Pesquisa Multidisciplinar); através da submissão das contribuições de analítica visual para revistas, de forma a obter avaliações de revisores especialistas.

6 Outros apoios

Projeto FAPESP, infraestrutura do ICMC/USP, grupo VICG, GSoC. NEXOS e labMacambira.sf.net para usuários e para realizar experimentos de interação nas próprias redes sociais.