

Mineração de Perfis Sociais em Redes Temporais

Jeancarlo C. Leão, Michele A. Brandão, Pedro O. S. Vaz de Melo,
Alberto H. F. Laender

Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG

{jcl,micheleabrandao,olmo,laender}@dcc.ufmg.br

Abstract. *The number and types of relationship that members of a social network establish suggest the profile that each individual develops over time. In this paper, we propose a mining method capable of identifying and recovering social profiles from topological aspects in temporal networks. The application of this method allows the identification of persistent profiles and explicit patterns that suggest relationship models in social networks.*

Resumo. *A quantidade e os tipos de relacionamento que os membros de uma rede social estabelecem sugerem o perfil que cada indivíduo desenvolve ao longo do tempo. Neste artigo, propomos um método de mineração capaz de identificar e recuperar perfis sociais a partir de aspectos topológicos em redes temporais. A aplicação desse método permite identificar perfis persistentes e explicitar padrões que sugerem modelos de relacionamento em redes sociais.*

1. Introdução

Uma rede social pode ser definida como uma estrutura composta de entidades em que o conjunto de interações entre os pares caracterizam algum tipo de relação. Tais redes podem ser modeladas por meio de um grafo, no qual os vértices representam as entidades e as arestas as interações ocorridas entre elas ao longo do tempo. Apesar do grande volume de dados existentes sobre redes sociais, eles geralmente são subutilizados em razão do pouco aproveitamento de suas propriedades temporais. Embora seja um desafio considerar a dimensão temporal, a topologia de uma rede social pode ser analisada em intervalos de tempo. Tal análise revela como cada entidade se relaciona, possibilitando determinar a sua tendência de interação ou perfis de relacionamento social. Nesse contexto, as informações topológicas e temporais de uma rede permitem responder à questão fundamental sobre qual seria o perfil social de seus nodos.

O estudo sobre perfis sociais é motivado pela escassez de abordagens para caracterização dos membros de uma rede social considerando apenas os seus aspectos topológicos e temporais. A obtenção de perfis sociais contribui para a quantificação do seu capital social¹, bem como permite caracterizar e distinguir diferentes redes sociais. Assim, são contribuições deste trabalho duas técnicas de mineração de redes sociais que atuam sobre classes de relações para obter o perfil social de relacionamento de cada um de seus

¹O capital social de uma rede é definido como sendo os recursos a ela incorporados que podem ser utilizados para obter melhorias como aumento do fluxo de informação, identificação de posições estratégicas na rede e elevação da credencial social de seus indivíduos [Lin 1999].

membros, e que foram avaliadas em redes sociais reais com diferentes características e de domínios distintos.

O restante deste trabalho está organizado da seguinte forma. A Seção 2 descreve trabalhos relacionados. A Seção 3 apresenta a metodologia adotada e uma breve caracterização das redes sociais utilizadas, introduzindo também a técnica adotada para classificar os seus relacionamentos. A seguir, a Seção 4 descreve as técnicas propostas para obtenção de perfis a partir das relações classificadas e a Seção 5 discute os resultados experimentais obtidos. Finalmente, a Seção 6 apresenta conclusões e comenta sobre trabalhos futuros.

2. Trabalhos Relacionados

Propriedades e padrões topológicos podem ser explorados para medir e distinguir tipos de relacionamento em redes sociais. Nesse contexto, David et al. [2010] analisam as funções que as arestas de uma rede desempenham em sua estrutura. Além disso, Granovetter [1973] define a força dos relacionamentos sociais a partir da estrutura topológica de uma rede. Tal conceito é utilizado por Alves et al. [2016] para medir a força dos relacionamentos em redes sociais de desenvolvedores e por Leão et al. [2017] para filtrar ruído informacional causado pela aleatoriedade em relações sociais.

O perfil social dos indivíduos de uma rede pode ser aplicado na obtenção de capital social mesmo quando os atributos desses indivíduos não estão acessíveis. Estudos em diferentes áreas propõem o uso das propriedades topológicas em aplicações distintas. Burt et al. [1995] argumentam que é útil se concentrar no padrão de relações entre pessoas. Podolny & Baron [1977] complementam o argumento de Burt et al. [1995] e apoiam-se também na teoria do capital social em que redes com alto fechamento (*closure*) e coerência (*structural holes*) são propícias à criação de uma identidade social que explica as expectativas por trás do seu papel.

Este trabalho baseia-se na força dos laços definida por Granovetter [1973] que pode ser obtida em redes temporais utilizando-se a técnica de classificação de relações proposta por Vaz de Melo et al. [2015]. Para isso, perfis sociais são definidos por meio da representatividade e importância dos relacionamentos para responder ao questionamento deixado por David et al. [2010] em relação aos papéis que diferentes nós desempenham na rede. Dessa forma, este trabalho propõe uma abordagem conjunta de meios já utilizados para estudo do capital social [Cohen et al. 2014; David et al. 2010] e revela informações ainda não observadas na literatura sobre os perfis sociais.

3. Metodologia

Dado um grafo não direcionado $G = (V, E)$, no qual $V = \{v_1, \dots, v_n\}$ é o conjunto de vértices e $E = \{e_1, \dots, e_m\}$ é o conjunto de arestas que representam as interações entre dois vértices, cada grafo temporal $G_t(V_t, E_t)$ em G representa a agregação das interações em períodos discretos de tempo t . Assim, para um dado valor de t , V_t inclui todos os vértices que interagem no período de tempo t . Analogamente, as arestas do conjunto E_t representam as interações entre os pares de vértices (v_i, v_j) durante o período de tempo t .

Esse modelo de grafo temporal foi aplicado a redes de colaboração científica derivadas de dados da APS, DBLP e PubMed², a redes *WiFi* de mobilidade em *campi*

²APS (<http://www.aps.org/>), PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>) e DBLP (<http://dblp.org/>)

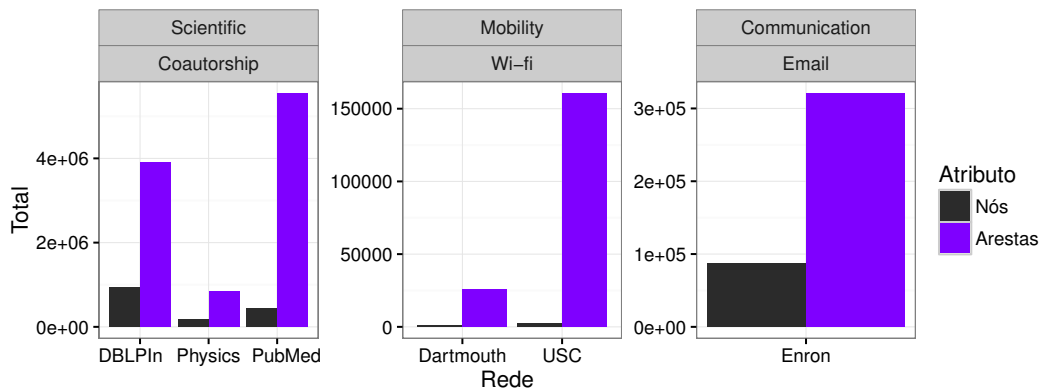


Figura 1: Características das redes sociais temporais.

Tabela 1: Classe atribuída quando as métricas estão acima (+) ou abaixo (-) do limiar l .
Fonte (adaptado): [Vaz de Melo et al. 2015].

Métrica comparada com limiar l	Classe <i>Amizade</i> (Forte e Persistente)	Classe <i>Conhecido</i> (Forte e Rara)	Classe <i>Ponte</i> (Fraca e Persistente)	Classe <i>Aleatório</i> (Fraca e Rara)
<i>Neighborhood Overlap (NO)</i>	+	+	-	-
<i>Persistence (P)</i>	+	-	+	-

universitários (Dartmouth e USC) e a uma rede de comunicação de *e-mails* entre funcionários da Enron³ entre 1999 e 2003. A Figura 1 apresenta as características dessas redes, onde os nós representam, respectivamente, pesquisadores, usuários do sistema e funcionários, e as arestas coautorias, pontos de acesso de rede sem fio e *e-mails* individuais.

O problema de classificação de relações em redes sociais consiste em atribuir um rótulo do conjunto $R = \{r_1, r_2, \dots, r_k\}$ a cada par de vértices $r(v_i, v_j)$. Assim, são considerados aspectos topológicos e temporais para determinar qual rótulo será atribuído à relação [Vaz de Melo et al. 2015]. De fato, estudos sociológicos revelam que a topologia em que pares (i, j) de indivíduos estão envolvidos sugere o tipo ou classe de relação entre eles [David et al. 2010; Granovetter 1973]. Neste trabalho, a classificação é obtida pela métrica *neighborhood overlap (NO)* [David et al. 2010] que considera os N vizinhos de um nodo (Equação 1) e pela métrica de persistência P dos relacionamentos que é dada pelo somatório dos intervalos de tempo em que ocorrem alguma interação entre os pares i e j (Equação 2).

$$NO(i, j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j| - 2} \quad (1) \quad P(i, j) = \frac{1}{t} \sum_{k=1}^t [(i, j) \in \varepsilon_k] \quad (2)$$

Para classificar as relações foi utilizado o algoritmo *RECAST (Random Relationship Classifier Strategy)* que é uma técnica de classificação de arestas em redes temporais [Vaz de Melo et al. 2015]. O *RECAST* baseia-se nas duas características que estão sempre presentes nas relações sociais: a persistência (P) das interações e o compartilhamento de vizinhos entre pares (NO) para atribuir a cada relação uma das classes, conforme comparação dos valores dessas métricas com um limiar l (Tabela 1).

³Enron email dataset: <https://www.cs.cmu.edu/~./enron/>

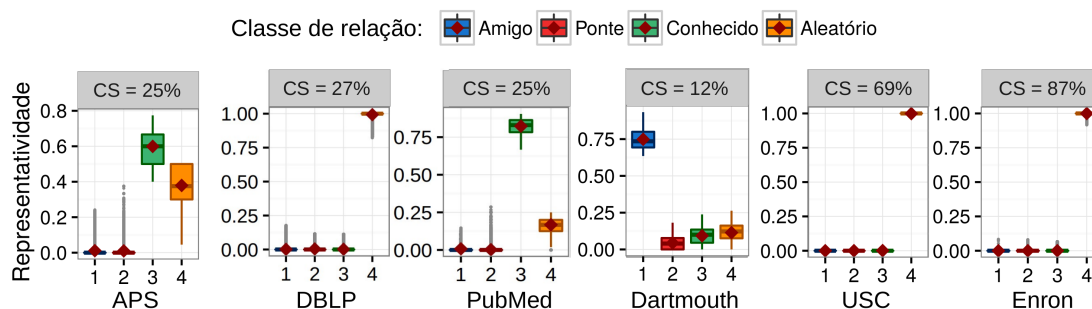


Figura 2: Perfis que caracterizam o maior grupo de indivíduos em cada rede.

4. Técnicas de Mineração de Perfis

O perfil social de um dado nodo de uma rede é caracterizado pela quantidade de relações que possui em cada classe e pode ser interpretado como a probabilidade que o nodo tem de estabelecer cada tipo de relação social. Dado o conjunto de todos os relacionamentos $R_i = \{r_1, r_2, \dots, r_n\}$ de um indivíduo i e sendo $R_{ic} \in R_i$ um subconjunto contendo todas as suas relações da classe c , definimos a representatividade das relações da classe c para o perfil do indivíduo i pela equação $Rep(i, c) = |R_{ic}|/|R_i|$. A importância $Imp(i, c)$ do conjunto de relacionamentos R_{ic} é relativa ao nodo j que possui maior número de relacionamentos da classe c na rede. Essa importância é obtida pela equação $Imp(i, c) = |R_{ic}|/|R_{jc}|$. Finalmente, o perfil de um indivíduo i é o conjunto formado pelas representatividades das quatro classes de relacionamentos.

Neste trabalho, propomos duas técnicas de recuperação de perfis. A primeira técnica permite identificar perfis existentes na rede através do agrupamento de seus nós pela representatividade de seus relacionamentos calculada por $Rep(i, c)$. A segunda técnica proposta se baseia no uso da métrica de importância de relacionamentos ($Imp(i, c)$) para obtenção de uma ordenação por classe para cada membro da rede. Assim, torna-se possível recuperar perfis relevantes a partir de uma consulta onde são especificados os níveis de importância de cada classe de relacionamento. Por exemplo, para elaborar uma consulta Q_1 que recupere os perfis com relacionamentos mais importantes da classe c , basta especificar um nível de importância alto para a classe c e baixo para as demais classes.

5. Resultados Experimentais

Os principais resultados deste trabalho são apresentados em dois segmentos de aplicação das técnicas propostas para mineração de perfis sociais: mineração de perfis similares e recuperação dos perfis com os relacionamentos mais importantes da rede. Em cada uma dessas aplicações, são analisadas também as características das redes utilizadas.

Mineração de Perfis Sociais Similares. A técnica de mineração de perfis sociais similares objetiva identificar os principais perfis sociais assumidos por membros de redes sociais com base nas classes de relacionamentos que possuem. Para isso é considerada a semelhança na representatividade de seus relacionamentos em cada classe.

Assim, para cada rede foram obtidos 10 agrupamentos de perfis dos quais os maiores (aqueles que possuem maior percentual CS de indivíduos da rede) são apresentados na Figura 2. É possível observar que nas redes DBLP, USC e Enron os maiores perfis

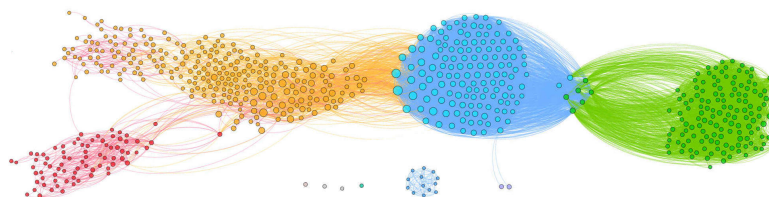


Figura 3: Subgrafo da rede PubMed com os indivíduos com perfis mais extremos.

possuem na maioria de suas relações interações raras ($P < l$) e pouco similares ($NO < l$), sendo, portanto, classificados como *aleatórios*. Por outro lado, o maior perfil na rede Dartmouth é extremamente social, pois seus membros mantêm, em sua maioria, relações regulares com pessoas com quem compartilham vizinhos na rede ($P > l$ e $NO > l$). As redes PubMed e APS, por sua vez, possuem um perfil com proporção considerável de relacionamentos da classe *conhecido*. Isto conclui que o maior grupo dessas redes mantém relações de coautoria pouco frequentes com um número relativamente grande dos coautores de seus coautores. Grupos que mantêm relações predominantemente de *ponte* são raros em todas as redes utilizadas neste trabalho a ponto de inviabilizar a formação de um grupo próprio utilizando o algoritmo *K-means* com $k = 10$. Isso confirma que pontes são raras em redes reais com propriedades de *Mundo Pequeno* [David et al. 2010].

Recuperação de Perfis Sociais Relevantes. Esta abordagem permite recuperar membros da rede com perfis relevantes para uma dada consulta definida pelos níveis de importância de cada classe de relacionamento. Por exemplo, ao consultar níveis de importância extremos, são selecionados membros com a importância mais expressiva na rede em cada classe. Na Figura 3, o resultado dessa consulta sobre a rede PubMed é apresentado como um subgrafo com grupos de perfis bem definidos e persistentes. Ou seja, para cada classe de relacionamento, os nós presentes nesse subgrafo são os que possuem maior importância em suas relações de amizade (em azul), com conhecido (verde), de ponte (vermelho) ou aleatórias (laranja) em toda a rede. Os atributos profissionais de alguns dos membros das redes de colaboração científica foram inspecionados individualmente revelando grupos de trabalho com padrões bem definidos para as diferentes classes de relacionamentos. Grupos com membros de mesmo perfil apresentam características topológicas similares em densidade, grau e coeficiente de agrupamento mesmo considerando sua participação em toda a rede. Ademais, é possível observar na Figura 3 que, no maior componente conectado, existe um grande grupo de membros altamente socializados (em azul), pois são persistentes e interagem muito e predominantemente com os demais membros desse grupo. Esses membros possuem características topológicas⁴ muito semelhantes entre si sugerindo forte influência social ou homofilia [David et al. 2010].

6. Conclusões

A contribuição fundamental deste trabalho está na modelagem da representatividade e da importância dos relacionamentos sociais. Por meio da aplicação dessas métricas foram concebidas técnicas para obtenção de perfis sociais que foram analisados em seis redes sociais reais. Dentre esses perfis, foram apresentados os que pertencem ao maior nú-

⁴As métricas topológicas locais foram medidas na rede completa. Dentre elas estão: centralidade, grau, coeficiente de clusterização e classe de modularidade.

mero de indivíduos, demonstrando diferenças no tipo de relacionamento que prevalece em cada domínio. Ademais, observamos que em redes de colaboração científica o grupo com maior número de indivíduos possui perfis diferentes, o que permite caracterizar e distinguir essas redes. Além disso, a técnica de recuperação de perfis relevantes permite elaborar consultas e extrair da rede o conjunto de membros com um determinado perfil de relacionamento. Os resultados também mostram como perfis com características mais extremas também podem ser obtidos para cada uma das classes de relacionamentos. As propriedades dos membros das redes, como a sua função profissional, grupo de trabalho ou área de atuação, foram analisadas permitindo relacionar os papéis reais desses indivíduos com o perfil obtido através de características puramente topológicas. Em trabalhos futuros, pretendemos verificar se os padrões dos relacionamentos e as similaridades topológicas observados em grupos de indivíduos com perfis extremos e similares podem ser explicados pela homofilia ou pela influência social.

Agradecimentos. Este trabalho é apoiado pelos projetos InWeb (processo MCT/CNPq 573871/2008-6) e MASWeb (processo FAPEMIG/PRONEX APQ-01400-14), e por auxílios individuais de pesquisa concedidos aos autores pelo CNPq, Fapemig e CAPES. Particularmente, o primeiro autor agradece ao Instituto Federal do Norte de Minas Gerais - IFNMG, pela concessão de uma bolsa de estudos no âmbito do Programa de Bolsas para Qualificação de Servidores (PBQS).

Referências

- Alves, G. B., Brandão, M. A., Santana, D. M., da Silva, A. P. C., and Moro, M. M. (2016). The Strength of Social Coding Collaboration on GitHub. In *Anais do 31º Simpósio Brasileiro de Banco de Dados*, pages 247–252, Salvador, Bahia, Brasil.
- Burt, R. S. (1995). *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge.
- Cohen, E., Delling, D., Pajor, T., and Werneck, R. F. (2014). Computing classic closeness centrality, at scale. In *Proceedings of the Second ACM Conference on Online Social Networks*, pages 37–50, Dublin, Ireland.
- David, E., Jon, K., Easley, D., and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA.
- Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.
- Leão, J. C., Brandão, M. A., de Melo, P. O. V., and Laender, A. H. F. (2017). Classificação de relações sociais para melhorar a detecção de comunidades. In *Proceedings of the VI Brazilian Workshop on Social Network Analysis and Mining*, São Paulo, SP, Brazil.
- Lin, N. (1999). Building a network theory of social capital. *Connections*, 22(1):28–51.
- Podolny, J. M. and Baron, J. N. (1997). Resources and relationships: Social networks and mobility in the workplace. *American Sociological Review*, 62:673–693.
- Vaz de Melo, P. O. S., Viana, A. C., Fiore, M., Jaffrès-Runser, K., Mouël, F. L., Loureiro, A. A. F., Addepalli, L., and Guangshuo, C. (2015). RECAST: Telling Apart Social and Random Relationships in Dynamic Networks. *Performance Evaluation*, 87:19–36.