Influência do Encaminhamento de Mensagens na Topologia de Redes Sociais

Samuel da Costa Alves Basilio, Gabriel de Oliveira Machado

[1]Centro Federal de Educação Tecnológica de Minas Gerais - CEFET MG, Unidade Leopoldina

Resumo. *Seja uma conversa entre amigos ou uma campanha publicitária, todo processo de comunicação dos usuários de redes sociais contém informações relevantes, e através do estudo dessas formas de comunicação é possível encontrar padrões que podem ser utilizados para diferentes fins, como marketing, disseminação de informações, formação de opiniões entre outros objetivos que possam ser alcançados pela comunicação com as massas. O objetivo deste trabalho é caracterizar uma rede social durante um espaço de tempo de uma forma dinâmica, observando separadamente a topologia da rede em pequenos espaços de tempo analisando como a troca e o encaminhamento de mensagens pode alterar a topologia da rede, considerando-se grupos relativamente pequenos de usuários, observando como estes padrões se alteram de acordo com o aumento do número de usuários observados.*

1. Introdução

O uso de redes sociais online não para de aumentar[Kwak et al. 2010]. Por exemplo, no terceiro bimestre de 2015 o Twitter relatou um total de 320 milhões de usuários mensais ativos[Twitter 2015]. Já o Facebook apresentou relatórios mostrando uma média de 894 milhões de usuários por dia e mais de 1.5 bilhões de usuários mensais ativos[Facebook 2015][Gjoka et al. 2010]. Com o aumento do número de usuários utilizando as redes sociais, existe também um crescimento do número de informações que acabam transitam nestas redes. Este tipo de informação pode trazer benefícios para aqueles que a detém, seja este benefício intelectual ou financeiro, como ocorre tradicionalmente com as propagandas que são direcionadas à um determinado público-alvo, de acordo com dados coletados dos usuários. Desta forma, as redes sociais passaram a despertar o interesse de pesquisadores de diferentes áreas do conhecimento, em especial os de computação, pois através desta é possível obter informações, analisa-las e a partir desta análise inferir e prever o comportamento destas redes quanto a vários aspectos, como, por exemplo, como as informações se espalham na rede ou ainda, como e quando ocorrem novas ligações entre usuários[Hutto et al. 2013][Khrabrov and Cybenko 2010].

O objetivo deste trabalho é observar o comportamento de amostras de diferentes tamanhos da rede do Twitter em diversos instantes no tempo, analisando sua dinamicidade, isto é, como ocorrem as alterações na topologia da rede no tempo, como a troca e o encaminhamento de mensagens nesse período influenciam essas mudanças e como a observação de diferentes tamanhos de amostras altera os resultados obtidos.

2. Metodologia de Captura e Análise dos Dados

Para o realizar deste estudo dois tipos de dados são necessários: A topologia da rede social e as mensagens trocadas pelos usuários da rede. A topologia da rede é obtida observando

os relacionamentos entre os usuários. Na maioria das redes sociais online pode-se representar esta topologia por um grafo direcionado, onde cada usuário é um nó do grafo e os relacionamentos são suas arestas. Com isso temos a topologia da rede representada pelo grafo G, onde $G = (V, E)$, com $V = \{u_x | u_x$ é um usuário da rede social$\}$ e $E = \{(u_i, u_j) | (u_i, u_j)$ é a relação existentes entre quaisquer usuários u_i e $u_j\}$. Em redes onde a amizade entre dois indivíduos é sempre recíproca pode-se até mesmo, dependendo do objetivo, simplificar esta estrutura e utilizar um grafo não-orientado. Chamaremos de instante inicial $t0$, o momento onde a rede é observada pela primeira vez e G_{t_0} o grafo que representa o estado da rede nesse momento. A partir deste ponto, é estabelecido um intervalo de tempo Δt, que é o tempo entre cada ponto subsequente de observação do estado da rede. Sendo assim, a rede será novamente observada no instante $t_1 = t_0 + \Delta t$, e novamente em um instante $t_2 = t_1 + \Delta t$, e assim sucessivamente, generalizado $t_n = t_{n-1} + \Delta t$ para $n > 0$, até que se obtenha o número n de amostras necessárias para a análise, sendo cada amostra representada por um grafo G_{tn}.

Simultaneamente à observação das mudanças do estado da rede também são observadas as mesagens enviadas por cada usuário u. Chamaremos de m_y^x uma mensagem enviada por um usuário u_x e vista por um usuário u_y. Pretende-se ao final do processo de captura e análise dos dados calcular uma métrica que corresponda à influência do encaminhamento de mensagem sobre a topologia da rede. Para isso, é preciso calcular a quantidade total de mudanças na topologia da rede D_{n+1}^n(**D**iferença), ocorridas entre dois instantes de observação t_n e t_{n+1}. Também é preciso calcular a quantidade de mudanças possivelmente influênciadas por uma troca de mensagem, ocorridas entre dois instantes de observação t_n e t_{n+1}, onde um usuário u_A que segue um usuário u_B, passou a seguir um usuário u_C, seguido por u_B depois de ter recebido uma mensagem de u_C encaminhada por u_B (Figura 1). À essa métrica chamamos de DI_{n+1}^n (**D**iferença **I**nfluênciada). Tanto D como DI são calculados comparando os grafos G_{tn} e G_{tn+1}. Sempre que uma aresta é criada ou removida somamos um valor a D e sempre que uma uma aresta é criada ou removida e a observação das mensagem indica um encaminhamento somamos um valor a DI. À influência do reencaminhamento de mensagens na rede entre t_n e t_{n+1} chamamos de I_{n+1}^n e calculamos $I_{n+1}^n = \frac{D_{n+1}^n}{DI_{n+1}^n}$.

Figure 1. Mudança na topologia da rede

3. Estudo de caso:Twitter

A captura de dados nas grandes redes sociais é o maior desafio desse tipo de projeto. No geral as grandes redes sociais não disponibilizam abertamente acesso a sua base para coleta dos dados necessários à esse tipo de pesquisa. Esse fato limita bastante os tipos de análises que podem ser realizadas. Em especial a troca de mensagens é quase sempre privada e as mensagens só podem ser vistas pelos usuários que estão se relacionando.

Essa restrição é menos rigorosa no Twitter[Twitter 2016a], por isso essa foi a rede social escolhida para a aplicação do presente estudo. Outra questão prática é que, mesmo com um acesso privilegiado à base de dados da rede social, analisar a rede por completo é computacionalmente muito difícil. Dessa forma adicionamos uma restrição no conjunto de usuários, observando apenas uma amostra do total de usuários e extrapolando os resultados obtidos para toda a rede.

Para que a coleta de dados seja feita no Twitter foi necessária a criação de uma conta de desenvolvedor[Twitter 2016b]. Somente com a criação de uma conta desse tipo é possível desenvolver aplicações que fazem a autenticação com a rede, autenticação que é necessária à uma aplicação que pretende acessar os dados da rede social. Porém mesmo com uma contade de desenvolvedor o acesso à API (*Application Programming Interface*) do Twitter é limitado para cada conta a 15 requisições por janela de tempo, sendo que cada janela de tempo dura 15 minutos. Desse forma, fora a conta do desenvolvedor, necessária para criar uma aplicação autorizada a acessar a API, foi necessário a criação de várias outras para que quando uma conta atingisse o limite de requisições uma outra conta pudesse autenticar a aplicação e continuar a captura dos dados.

Até o presente momento observamos um total de 50000 usuários. A escolha dos usuários que seriam analisados foi feita de forma aleatória, restringindo apenas o fator geográfico para que a ligação entre os usuários observados fosse mais provável. As informações capturadas na rede observada são: número de relacionamentos totais em cada instante; alteração no número de relacionamentos entre dois instantes; número de novos relacionamentos entre dois instantes; número de relacionamentos desfeitos entre dois instantes; número de mensagens trocadas entre dois instante (a partir do início do teste); número de mensagens encaminhadas entre dois instantes (a partir do início do teste). O período Δt escolhido foi de 24 horas, e foram coletadas 7 amostras. Sendo assim, tivemos um mapeamento da rede por dia em um período de uma semana. A tabela 1 apresenta os dados coletados.

Table 1. Quantificação dos dados coletados

	$t1$	$t2$	$t3$	$t4$	$t5$	$t6$	$t7$
Total de relacionamentos por tempo	799	1005	1043	1070	1081	1107	1127
Total de novos Follows	0	212	52	32	30	28	26
Total de Unfollows	0	6	14	5	19	2	6
Mensagens enviadas	0	26641	46764	67839	89342	111364	136196
Retweets	0	6	14	29	38	46	54

Após a realização de alguns testes, foi possível obter alguns resultados quanto a aspectos básicos com relação à rede, tais como o número de mensagens reencaminhadas em relação ao número de mensagens totais da rede, o número de ligações estabelecidas dentro da rede, dentre outros. A seguir, veremos os resultados de dois testes iniciais realizados com a utilização da aplicação desenvolvida.

3.1. CONECTIVIDADE DA REDE

O objetivo deste primeiro teste foi estabelecer uma relação entre o crescimento da amostra observada e o aumento das ligações entre os usuários, isto é, o aumento da conectividade.

Para a realização de tal teste, foi estabelecido que as medições seriam feitas nos valores de mil, 5 mil, 10 mil, 15 mil, e assim sucessivamente, até os 50 mil usuários observados. Para cada valor deste, foi capturada a rede e então estabelecida a conectividade da rede, avaliando o número de relações entre usuários da própria rede. Os resultados para este teste podem ser observados no gráfico apresentado na Figura 2.

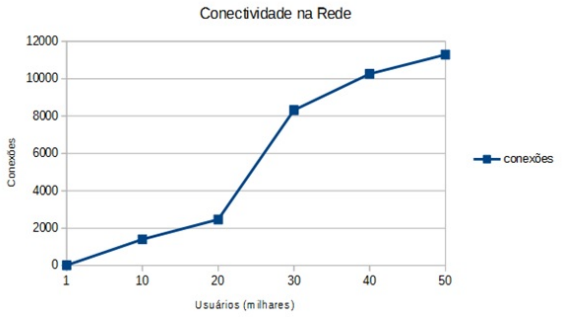

Figure 2. Conectividade entre os usuários da rede

Como podemos ver, o número de relações entre os usuários cresce de acordo com que o número de usuários aumenta, como era esperado. Podemos observar ainda que, para os últimos valores de usuários, a rede apresenta um decréscimo no aumento da conectividade. Esta diminuição pode ser resultado da coleta de usuários que possuem menos conectividade entre si, uma vez de o número de usuários coletados é maior.

É interessante ressaltar a importância deste teste na realização das análises subsequentes, uma vez que fica conhecido, de forma aproximada, qual a conectividade esperada para uma rede que possua um número de usuários pertencente à faixa descrita anteriormente. Com a realização de outros testes, foi possível comprovar a veracidade de, pelo menos, uma das informações obtidas, como veremos a seguir.

3.2. CONECTIVIDADE EM FUNÇÃO DO TEMPO

Como a proposta desse trabalho é caracterizar uma rede de forma dinâmica, fica clara a realização de uma análise de variação do estado da rede de acordo com o tempo. Para a realização de tal teste, foi empregada a metodologia descrita anteriormente, isto é, foram coletados dados em vários instantes para avaliar a variação da rede entre eles. Assim, como frequência para a coleta de dados, foi estabelecido o período de 24 horas de intervalo, com uma quantidade total de 7 coletas do estado da rede, resultando em uma semana de observação. As informações sobre os dados coletados foram exibidas na Tabela 1.

O primeiro dos resultados obtidos com estes testes pode ser observado no gráfico representado na Figura 3, que indica o número de relacionamentos em uma rede de 10 mil usuários, de acordo com os dias de análise.

Podemos observar que, à medida que o teste avança, o aumento do número de conexões entre os usuários diminui, indicando que, provavelmente, este número tenda a zero. Assim, podemos concluir que uma rede de determinado tamanho, fixo no tempo, tende a estabilizar no que diz respeito a sua topologia, isto é, a frequência com que os usuários criam novas relações entre si reduz à medida que o tempo passa. É observado também um ligeiro aumento no número de novas ligações nos dias 6 e 7. Este aumento é,

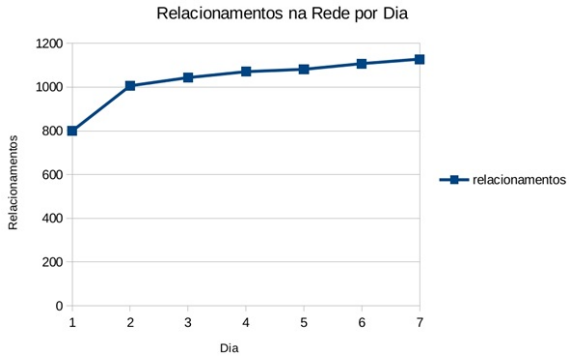

Figure 3. Conexões em uma rede durante o período de análise

provavelmente, ocasionado por uma maior presença dos usuários na rede, o que perfeitamente normal, uma vez que tais dias representam o final de semana, onde os usuários têm mais tempo para acessar suas contas no Twitter.

No entanto, o gráfico apresentado anteriormente apenas indica o número de conexões existentes na rede em determinado instante. Para que saibamos com exatidão qual o número de novas interações entre os usuários entre um instante e outro, devemos considerar o número de interações que deixaram de existir, isto é, o número de unfollows que ocorreram no dado período. Para tanto, a aplicação também é capaz de identificar, separadamente, o número de novas ligações feitas e de ligações que deixaram de existir através da análise dos resultados da coleta de dados.

Nesta mesma análise, foram capturadas as mensagens enviadas pelos usuários no período analisado, salvando tais mensagens de forma conjunta aos arquivos de estado da rede. Assim, com a análise dos dados obtidos, foi possível analisar a diferença no crescimento do número de mensagens reencaminhadas em relação ao número de mensagens totais enviadas, como mostra a Figura 4.

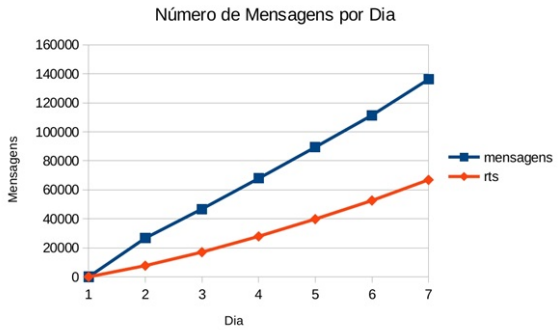

Figure 4. Mensagens por dia

É possível perceber que a curva de crescimento das mensagens reencaminhadas (RTs) demonstra um aumento maior a cada dia. Ao mesmo tempo, a curva que demonstra o número de mensagens enviadas apresenta uma curvatura muito mais suave, o que implica que o número de mensagens reencaminhadas aumenta de forma mais acelerada que o número total de mensagens.

4. Conclusão e Trabalhos Futuros

Como foi visto, o trabalho já apresenta resultados iniciais com relação a análise dinâmica da rede do Twitter, o que indica um bom funcionamento da metodologia aplicada, assim como das aplicações desenvolvidas para a realização efetiva desta análise. No entanto, é conhecido que vários aspectos ainda podem ser melhor desenvolvidos, tal como o estabelecimento da métrica para redes de outros tamanhos. No entanto, também são conhecidas as possíveis dificuldades para a realização de testes com uma escala muito maior do que a utilizada atualmente, principalmente devido à falta de acesso privilegiado à base de dados do Twitter. Disponibilizando apenas de acesso não-privilegiado, podem haver atrasos na coleta de dados, o que impossibilita a realização de testes que exigem um tempo pequeno de atraso na obtenção das informações. É ainda objetivada a determinação de um método pelo qual seja possível estimar métricas para uma rede de um tamanho para o qual ainda não tenham sido realizados análises. Sendo assim, os valores serão estimados com base em testes anteriores, em tamanhos reduzidos, que proporcionem uma estimativa para qualquer valor dado para o número de usuários.

References

Facebook (2015). Facebook q3 2015 earnings. http://investor.fb.com/.

Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2010). Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9.

Hutto, C., Yardi, S., and Gilbert, E. (2013). A longitudinal study of follow predictors on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 821–830, New York, NY, USA. ACM.

Khrabrov, A. and Cybenko, G. (2010). Discovering influence in communication networks using dynamic graph analysis. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 288–294.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

Twitter (2015). Twitter reports third quarter 2015 results. http://about.twitter.com/.

Twitter (2016a). Twitter. http://twitter.com/.

Twitter (2016b). Twitter developers. http://dev.twitter.com/.