mini:1

Capítulo 1

É uma questão de tempo! Extraindo Conhecimento de Redes Sociais Temporais

Fabíola S. F. Pereira, João Gama, Gina M. B. de Oliveira

Abstract

Data is structured as a network. And now? How to analyze it? Extracting knowledge from network data is not a simple task and requires the use of appropriate tools and techniques, especially in scenarios that take into account the volume and evolving aspects of the network. In this chapter it is considered that data has already been collected and is already structured as a network. The goal is to discuss techniques to analyze this network data, especially considering the time perspective. First, concepts related to problem definition, temporal networks and metrics for network analysis will be presented. Next, in a more practical aspect will be shown techniques of visualization and processing of temporal networks. In the end, three case studies with real data will be discussed, illustrating how network data knowledge extraction works from start to finish.

Resumo

Os dados estão estruturados na forma de rede. E agora? Como analisá-los? Extrair conhecimento desse tipo de dado não é uma tarefa simples e requer o uso de ferramentas e técnicas adequadas, especialmente em cenários que levam em conta o volume de dados e o aspecto temporal da rede. Neste capítulo considera-se que os dados já foram coletados e já estão estruturados em forma de rede e discute-se sobre técnicas para analisá-los, considerando especialmente a perspectiva temporal. Primeiro serão apresentados conceitos relacionados à definição do problema, redes temporais e métricas para análise de rede. Em seguida, em um aspecto mais prático serão mostradas técnicas de visualização e processamento de redes temporais. Ao final, três estudos de caso com dados reais serão discutidos, ilustrando do começo ao fim como funciona a extração de conhecimento de dados em rede.

1.1. Introdução

Redes sociais de amizade, redes de *hyperlinks*, de confiança e redes de co-autoria são exemplos de dados estruturados na forma de redes que representam entidades ligadas por alguma relação em comum. Análise de redes sociais é o campo de estudo que busca entender a estrutura e comportamento dessas redes, bem como as entidades que a ela pertencem [24]. Recentemente, houve um crescente interesse da comunidade de mineração de dados nesse campo de análise de redes sociais. A motivação básica é a demanda por explorar o conhecimento de grandes volumes de dados coletados, pertencentes ao comportamento social dos usuários em ambientes *online* [30].

Existe uma vasta literatura acerca de como coletar, pré-processar e modelar dados de mídias sociais em forma de redes [8], bem como acerca das principais métricas de centralidade [1]. Porém, ainda há muito a ser discutido em relação à análise da rede obtida. Por onde começar a análise de uma rede? E se ela for muito grande, como visualizá-la? Como apresentar os resultados obtidos? Quais as melhores técnicas para filtros e processamento? E para considerar a evolução temporal é melhor trabalhar com *snapshots*?

Neste capítulo então, considera-se que os dados já foram coletados e já estão estruturados em forma de rede e discute-se sobre técnicas para analisá-los, considerando especialmente a perspectiva temporal. É uma perspectiva pouco explorada e muito útil dentro do contexto de ciência de dados em rede.

1.1.1. Entendendo e Formalizando o Problema

Ao receber uma coleção de dados estruturados na forma de rede e a informação sobre o domínio que aquela rede representa (ex.: relações de amizade, trocas de e-mails, contatos visuais entre pessoas etc), o primeiro passo que um cientista de dados deve tomar é visualizar a rede recebida. A visualização pode ser feita com o auxílio de ferramentas (Seção 1.4) e, normalmente, sobre uma amostra obtida do todo.

Ao visualizar a rede, é possível rapidamente obter *insights* acerca dos dados em questão, tais como: quais características descrevem um nó e uma aresta, se a rede possui uma configuração parecida com algum modelo antes visto (esparsa, comunidades pequenas, bipartite), se possui informação temporal nas arestas e qual a granularidade dessa informação, se originalmente é dirigida, ou ainda, se originalmente possui diferentes tipos de nós. Note que a visualização de uma rede contempla não só a imagem estrutural, como também, os dados que a formam.

Uma vez observada a rede inicial, o segundo passo de uma análise é voltar ao problema em questão: qual é o problema que deseja-se resolver? Em alguns cenários o problema pode ser claramente solicitado, por exemplo, obter as comunidades da rede; em outros, o problema não está claro, e deve-se começar por uma análise exploratória que, futuramente, ajudará na elaboração de hipóteses para a definição do problema. Uma análise exploratória é, por exemplo, obter algumas medidas de influência, modularizar baseado em diferentes características dos nós, filtrar, observar nós *egos* (ego network [30]) etc. Dependendo do cenário, um análise exploratória já é suficiente para quem forneceu a rede. A entrega é, portanto, um conjunto de estatísticas que descrevem a rede em questão.

Um passo adiante da análise de exploratória é, finalmente, levar em consideração o aspecto temporal. Mesmo que o problema tenha claramente indicado que o tempo é fator fundamental, é importante passar pelas etapas básicas de conhecimento dos dados primeiro. Elas ajudam a perceber a real necessidade do tempo ser considerado. Em geral, problemas que consideram redes temporais estão relacionados à extração de conhecimento que leva em conta o fluxo natural de evolução da rede, ou o fluxo de informações que propagam na rede [28, 27, 23]. Por exemplo, é natural entender uma rede de contatos (aperto de mãos) entre pessoas em um contexto de transmissão de doenças. A Figura 1.1 sumariza as ideias até então apresentadas.

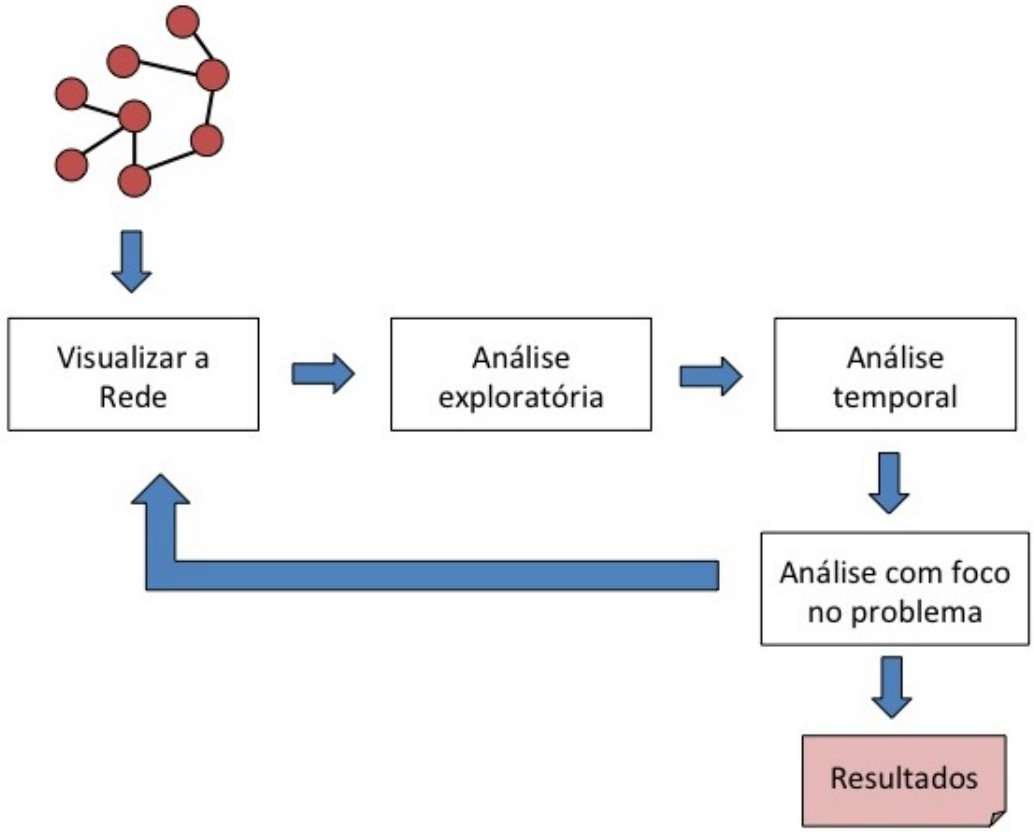

Figura 1.1. Passos para entender o problema da descoberta de conhecimento em redes.

É importante seguir tais passos para manter a visão crítica e conseguir discernir se uma rede que envolve a complexidade de uma análise temporal é realmente necessária. Ao longo do capítulo considera-se que o fluxo de descoberta de conhecimento em redes é respeitado e executado de maneira recorrente, até que se consiga respostas para um determinado problema.

1.1.2. Organização do capítulo

Na Seção 1.2 são apresentados os conceitos acerca de redes sociais temporais, bem como as principais métricas para analisá-las. Na Seção 1.3 são apresentadas as diferentes estratégias para processar redes evolutivas durante a análise. Tais estratégias variam entre processamento em blocos ou processamento de *streams* de redes. A Seção 1.4 é uma sín-

tese sobre estratégias de visualização de redes sociais temporais. A Seção 1.5 descreve três ferramentas principais para análise de redes. Na Seção 1.6 são apresentados três estudos de caso com bases de dados reais, destacando os processos utilizados na análise de cada rede para extração de conhecimento. Por fim, a Seção 1.7 apresenta as considerações finais.

1.2. Redes Sociais Temporais

Estruturas de redes representam relacionamentos entre entidades. Nas redes temporais, os tempos em que esses relacionamentos estão ativos são elementos explícitos da representação [11]. Um exemplo clássico de aplicação de uma rede temporal é o contágio de doenças através da proximidade física. Comumente, a propagação de organismos patogênicos ocorre através de um aperto de mão e uma rede temporal é a melhor estrutura para representar esse cenário. Redes sociais, tópico de interesse deste capítulo, também podem ser representadas como redes temporais, já que estão cada vez mais ubíquas e complexas em suas interações [10].

Existem várias definições na literatura que formalizam redes temporais (aqui, invariavelmente chamadas também de grafos temporais) [29, 17, 15, 11]. [15] definiu os grafos ordenados no tempo e [17] os chama de grafos que variam com o tempo, mas geralmente todas as definições representam um conjunto de arestas temporais e um conjunto de nós durante um intervalo de observação que leva em conta a ordem temporal em que aparecem (ou estão ativos).

Definição 1 (Rede temporal) *Uma rede temporal $G = (V, E)$ é um conjunto E de arestas registradas em meio a conjunto de nós V durante um intervalo de observação $[0, T]$. Uma aresta entre dois nós $u, v \in V$ é representada por uma quádrupla $e = (u, v, t, \delta t)$, onde $0 \leq t \leq T$ é o momento que a aresta surgiu e δt é sua duração. As arestas também são chamadas de contatos ou ligações.*

Essa definição é clássica para representar grafos de voos e redes de chamadas telefônicas, por exemplo. Mas existem extensões para a definição acima. Quando os contatos são instantâneos, $\delta t \to 0$, a rede temporal é definida como um grafo de sequência de contatos [11]. Esses grafos são usados para representar sistemas cuja duração do contato é menos importante (redes de e-mails, sexuais, redes de *likes* em redes sociais). Outra variação ao invés de definir redes temporais com arestas que não estão ativas sobre um conjunto de instantes, é defini-las sobre um conjunto de intervalos $e = (u, v, t_{init}, t_{end})$. Estes são os conhecidos grafos de intervalo, bons para modelagem de relacionamentos do tipo seguidor/seguido no Twitter [22]. De fato, grafos de intervalos podem ser transformados em grafos de sequência de contatos e, então, a maioria das técnicas de análise de redes pode ser empregada independente da modelagem utilizada.

Exemplo 1 *A Figura 1.2 ilustra duas redes temporais, considerando o contexto da rede social Twitter. A Figura 1.2(a) é um grafo de sequência de contatos, representando menções entre os usuários. Os nós são usuários e uma aresta (u, v, t) indica que u mencionou v em um tweet postado no tempo t*[1]*. Os instantes que ocorrem as interações estão descri-*

[1]Menção é um tweet que contém uma referência a outro usuário

tos próximos às arestas e a duração das interações é negligenciável. É possível perceber que os usuários A e B interagiram nos instantes 3, 6 e 11, os usuário B e C interagiram em 7 e 9 e assim por diante.

Agora, no mesmo contexto do Twitter, é possível considerar um grafo de intervalos na Figura 1.2(b), onde as arestas representam relações de seguidores/seguidos e os intervalos indicam que tais relações começaram em t_{init} e terminaram em t_{end}. Como exemplo, E começou a seguir F em 3 e deixou de segui-lo em 6.

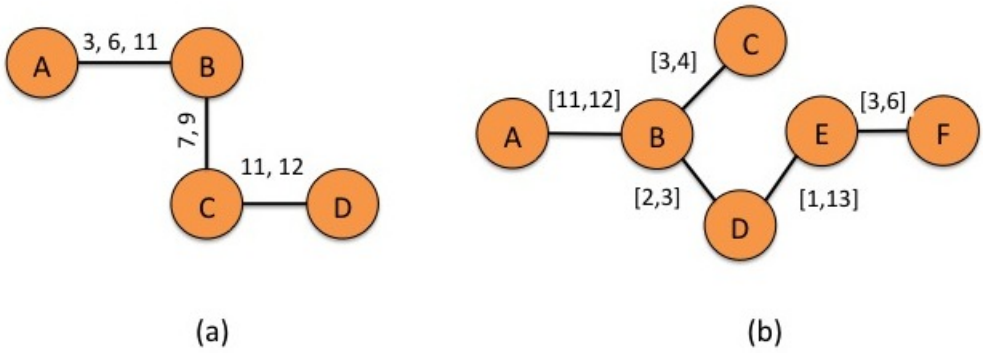

Figura 1.2. Redes temporais representadas como (a) sequência de contatos e (b) grafo de intervalos.

1.2.1. Métricas Temporais de Redes

A estrutura topológica de redes estáticas pode ser caracterizada por uma série de métricas [7, 26]. Em essência, tais medidas são baseadas em conexões entre nós vizinhos (tais como *degree* ou coeficiente de clusterização), ou entre grandes conjuntos de nós (caminhos, diâmetro e métricas de centralidade). Quando a dimensão tempo é incluída na rede, muitas dessas medidas precisam ser repensadas.

1.2.1.1. Caminhos em função do tempo

Em um grafo estático, um caminho é simplesmente uma sequência de arestas tais que uma aresta termina no nó onde a próxima aresta inicia (tal como o caminho A para B para C para D na Figura 1.2. Em um grafo temporal, caminhos são definidos como sequências de contatos com tempos crescentes que conectam um conjunto de nós – os caminhos em função do tempo [14]. Como exemplo, na Figura 1.2(b) existe um caminho em função do tempo de A para B (<(A,B,11)>, por exemplo), mas nenhum de A para D.

Uma diferença entre redes estáticas e temporais é que os caminhos não são transitivos. A existência de um caminho em função do tempo de i para j e de j para k não implica na existência de um caminho de i para k. Esse fato está relacionado a uma propriedade fundamental nos caminhos em função do tempo – um caminho de i para k via j existe somente se o primeiro contato entre j e k ocorreu depois de um contato em i e j.

Portanto, caminhos em função do tempo definem quais nós podem ser atingidos a partir de outros nós dentro de uma janela de observação $t \in [0,T]$. O conjunto de nós

que podem ser atingidos a partir de um nó i é chamado de conjunto de influência de i. No contexto de redes sociais, por exemplo, o conjunto de influência será atingido pelos *posts* de i.

A duração de um caminho em função do tempo é a diferença entre o último e o primeiro contato de um caminho [20]. Analogamente ao conceito de menores caminhos em redes estáticas que define a distância geodésica, em redes temporais existem os *caminhos mais rápidos em função do tempo*, que indicam caminhos com menor duração. No exemplo (Figura 1.2)(b) existem vários caminhos em função do tempo de B para F na janela de observação $[3,15]$: <(B,D,2),(D,E,4),(E,F,5)>, <(B,D,3),(D,E,4),(E,F,6)>, etc.. O caminho mais rápido em função do tempo possui duração 3.

1.2.1.2. Medidas de Centralidade Temporal

Na teoria de rede, diversas medidas de centralidade foram definidas para identificar o comportamento dos nós e arestas, muitas delas fundamentadas no conceito de menores caminhos. Trazendo para o cenário de redes temporais, a ideia é interpretar tais medidas utilizando os caminhos mais rápidos em função do tempo (ao invés de menores caminhos).

Closeness temporal. A métrica de centralidade closeness C_C [5] para redes estáticas é definida como

$$C_C(i) = \frac{N-1}{\sum_{j \neq i} d(i,j)} \tag{1}$$

onde $d(i,j)$ é a menor distância geodésica entre i e j, i.e. a centralidade closeness de um nó mede o inverso da menor distância total para todos os outros nós e é alta para aqueles nós mais próximos de todos os outros (centrais). Similarmente, em redes temporais, a ideia é medir o quão rápido um nó pode em média atingir os demais:

$$C_C(i,t) = \frac{N-1}{\sum_{j \neq i} \lambda_{i,t}(j)} \tag{2}$$

onde $\lambda_{i,t}(j)$ é a latência entre i e j, definida por $\lambda_{i,t}(j) = t - \phi_{i,t}(j)$, sendo $\phi_{i,t}(j)$ o instante mais recente antes de t em que a informação de j pode ter atingido i. A latência mede a idade da informação em um nó.

Betweenness temporal. A centralidade betweenness C_B [5] é também baseada em menores caminhos. Ela mede a fração entre o número de menores caminhos que passam pelo nó em questão em função do número total de menores caminhos entre cada par de nós na rede. Para redes estáticas, tal centralidade é formalmente definida por

$$C_B(i) = \frac{\sum_{i \neq j \neq k} v_i(j,k)}{\sum_{i \neq j \neq k} v(j,k)} \tag{3}$$

onde $v_i(j,k)$ é o número de menores caminhos entre j e k que passam por i, e $v(j,k)$ é o número total de menores caminhos entre j e k. Pensando no contexto de redes temporais, tem-se:

$$C_B(i,t) = \frac{\sum\limits_{i \neq j \neq k} w_{i,t}(j,k)}{\sum\limits_{i \neq j \neq k} w_t(j,k)} \tag{4}$$

onde $w_{i,t}(j,k)$ é o número de caminhos mais rápidos em função do tempo na janela de observação t entre j e k que passam por i e $w_t(j,k)$ é a quantidade total de caminhos mais rápidos em função do tempo.

Além das métricas mais populares de closeness e betweenness, em [11] são discutidas diversas outras métricas que se aplicam ao contexto temporal.

1.3. Análise de Redes Evolutivas

Quando o assunto são redes que evoluem ao longo do tempo, uma questão importante é discutir como processá-las. Por exemplo, em redes de e-mails as arestas são adicionadas a cada minuto, enquanto em redes de co-autoria as arestas surgem em escalas de semanas ou meses. Assim, mesmo considerando os aspectos de evolução e informação temporal, é necessário refletir se análises *offline* ou em tempo real são necessárias em determinado contexto.

1.3.1. Estratégias para Manipular a Evolução da Rede

Aqui estão sumarizadas diferentes estratégias de processamento de redes que evoluem ao longo do tempo. Essas estratégias estão relacionadas com a rede – ou amostras dela – que são consideradas durante a análise, impactando (i) no processo de ajuste do modelo, (ii) performance dos algoritmos e (iii) interpretação semântica que o analista está interessado. O termo *ordem temporal das arestas* se refere à ordem em que as arestas chegam na rede, ou seja, se uma rede temporal é considerada na análise, conforme discutido nas seções anteriores. Na Figura 1.3 está ilustrado um cenário em que as arestas chegam em fluxo (*edge stream*) que será usado para exemplificar as estratégias.

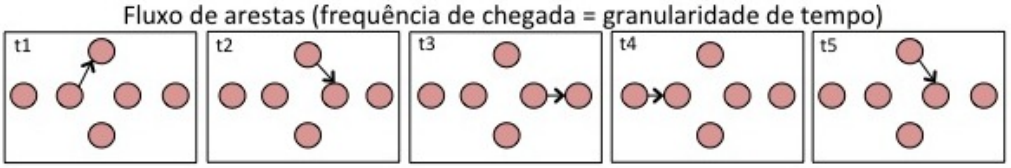

Figura 1.3. Fluxo de arestas

- *Redes que evoluem lentamente (slowly evolving networks).* A maioria dos trabalhos propostos na década passada considera essas estratégias quando processam redes [9]. Elas são intuitivas e diretas, conforme mostrado na Figura 1.4.

– *Processamento em blocos (batch processing).* Toda a rede é simplesmente processada considerando a ordem temporal das arestas. Algoritmos clássicos como Dijkstra são usados nesse cenário.

– *Snapshots.* A cada instante $t1, t2, ...$ um snapshot da rede é considerado. Aqui, a ordem temporal só faz sentido se a granularidade dos snapshots for maior que a granularidade da ordem de chegada das arestas.

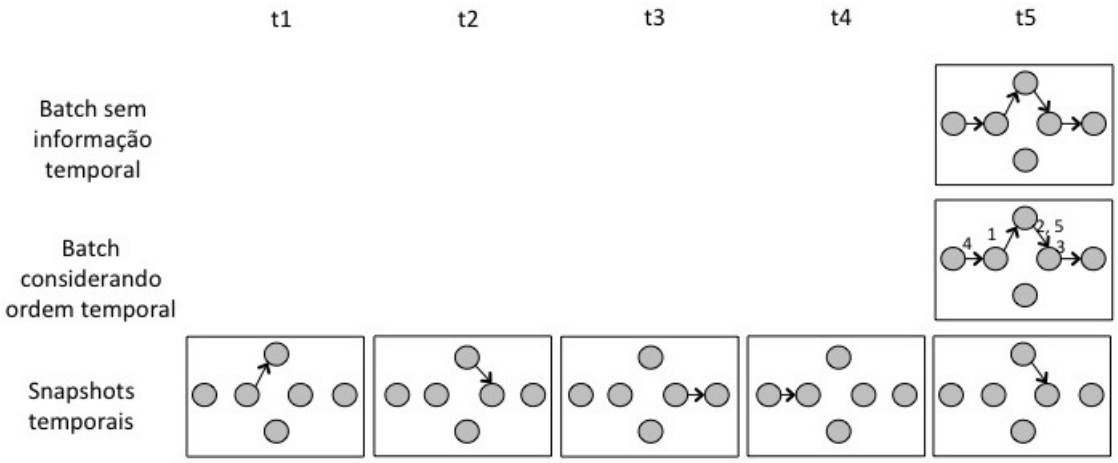

Figura 1.4. Estratégias para processamente de redes que evoluem lentamente considerando o fluxo de arestas apresentado na Figura 1.3

- *Redes stream.* Esse cenário é muito mais desafiador em termos de algoritmos devido a restrições computacionais e à incapacidade de carregar toda a rede no disco. O que varia é a abordagem de janelas e se a ordem temporal das arestas também é levada em conta dentro das janelas (Figura 1.5)). Independente da estratégia, em redes stream os algoritmos deve usar estruturas de dados que podem ser mantidas incrementalmente.

  – *Janela landmark.* Essa estratégia é boa quando deseja-se manter o histórico e as novas arestas que chegam são processadas considerando todo o grafo armazenado até então. Levar em conta a ordem temporal (redes temporais) é um cenário muito comum. Por exemplo, redes de contato entre pessoas são processadas utilizando essa estratégia num contexto de propagação de doenças [11].

  – *Janela deslizante (sliding window).* Na janela deslizante o passado recente da rede é suficiente. Essa é a estratégia base para algoritmos de amostragem com fator de esquecimento [2]. Não é usual considerar a ordem temporal dentro da janela.

  – *Janela de observação fixa.* Alguns trabalho processam a rede com uma janela de observação fixa [29], na qual apenas um determinado intervalo é interessante. Grafos de voos são processados usando essa estratégia, considerando redes temporais dentro da janela.

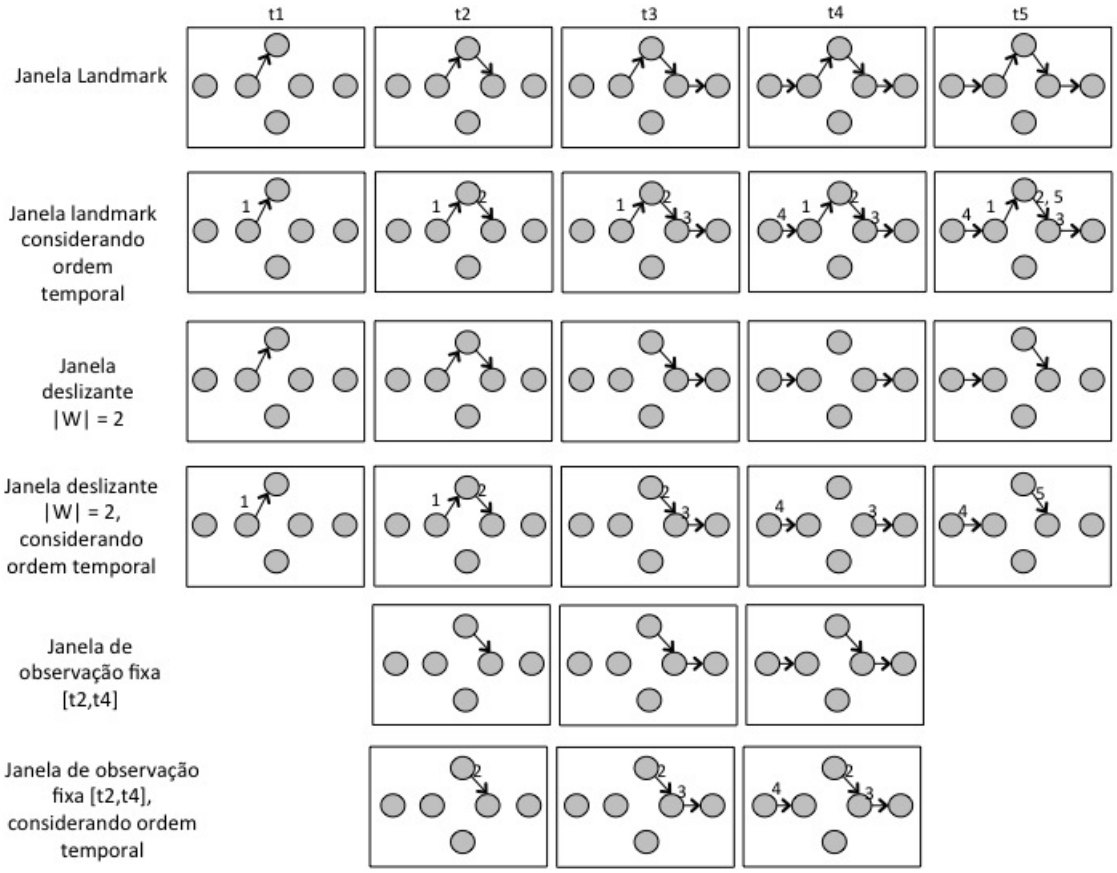

Figura 1.5. Estratégias para processamento de redes stream considerando o fluxo de arestas da Figura 1.3

1.3.2. Nivelando os conceitos

Não existe um consenso na literatura em relação aos termos utilizados para expressar redes que evoluem ao longo do tempo. Os conceitos mais frequentemente empregados são:

- Rede evolutiva (*evolving network, evolutionary network*) ou rede evolutiva no tempo (*time-evolving network*) ou rede dinâmica (*dynamic network*) ou rede com variação no tempo (*time-varying network*). Todos esses conceitos referem-se a redes que estão mudando, com nós aparecendo e desaparecendo, associando e desassociando uns com os outros à medida que o tempo passa.
- Rede temporal. Como detalhado na Seção 1.2, redes temporais são redes cuja ordem em que as arestas aparecem e desaparecem é levada em conta durante a análise – a ordem temporal.
- Rede *stream*. Esse termo está relacionado à maneira que uma rede é observada e processada, especialmente em cenários onde não existe a noção de começo e fim do fluxo de chegada dos dados.

1.4. Visualização de Redes

O uso de recursos visuais pode auxiliar na obtenção de *insights* durante o processo de análise. De fato, técnicas de visualização de redes têm sido muito exploradas atualmente [3, 16, 12, 18].

A visualização auxilia principalmente no processo inicial da análise de uma rede temporal. Por exemplo, considere que diante de uma rede na qual apenas o domínio é conhecido – rede temporal de contatos entre pessoas e possível transmissão de doenças. A Figura 1.6 é um exemplo de visualização que pode ser aplicada na rede. A partir da figura é possível obter os seguintes *insights*:

1. A pessoa D inicialmente é muito ativa, mas à medida que o tempo passa ela deixa de entrar em contato com as demais pessoas;
2. As pessoas B e C são muito ativas na rede;
3. A pessoa A tem baixo nível de contatos;
4. Se a pessoa A possuir o vírus da gripe na altura do tempo $t = 6$, apesar do seu baixo nível de interatividade, D nunca será infectado, porém B e C serão.

Se a visualização temporal, entretanto, não tivesse sido utilizada, provavelmente a rede de contatos entre pessoas seria representada como na Figura 1.7. Apenas olhando para a rede estática, não é possível extrair informações da sequência de contatos e, erroneamente, poderia-se concluir que se A possuir vírus da gripe, todos serão infectados.

Primordialmente, a visualização é, então, utilizada para obtenção de insights por parte do analista da rede. Porém, as ferramentas de visualização são acionadas também no

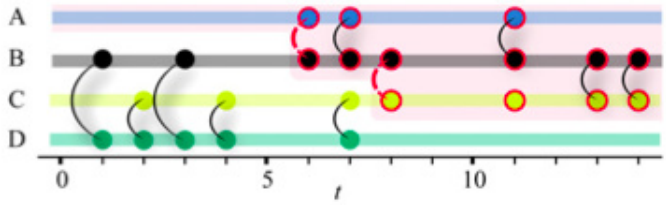

Figura 1.6. Obtida de [11]. Representação explícita da dimensão temporal da rede que ilustra uma sequência de contatos entre pessoas e possível transmissão de doença na rede.

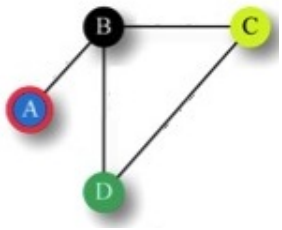

Figura 1.7. Adaptada de [11]. Rede estática de contatos entre pessoas.

momento da apresentação de resultados. Considere, por exemplo, que ao final da análise de uma rede temporal, queira-se destacar o padrão de evolução da centralidade de um nó. As técnicas de visualização buscam maneiras de representar tal nó (ou comunidades) em destaque no meio de uma infinidade de nós e arestas se cruzando.

1.5. Ferramentas para Análise

As principais ferramentas para análise e visualização de redes sociais temporais são enumeradas a seguir:

1. Gephi [4]. Ferramenta *open source* construída em Java, com arquitetura flexível, pronta para receber plugins. Essa flexibilidade tem tornado o Gephi a ferramenta mais popular para análise de redes dinâmicas. A Figura 1.8 ilustra uma rede social desenhada através dessa ferramenta e a Figura 1.9 destaca a funcionalidade temporal disponível no Gephi.

2. ORA [6]. A principal característica dessa ferramenta é a organização e disponibilização de funcionalidades, que estão dispostas com um nível maior de abstração. Por exemplo, ao invés de ter funcionalidades como calcula *closeness*, *betweenness*, ORA possui funcionalidades como *obtém mais influentes*, *obtém nós que são pontes* e assim por diante. É uma ferramenta cujo público-alvo não precisa ser técnico para conseguir manipulá-la, como por exemplo jornalistas. ORA suporta redes sociais padrões (Twitter, Facebook), redes organizacionais, geo-espaciais, *meta-networks*, redes dinâmicas entre outras. A Figura 1.10 é um *snapshot* da ferramenta.

3. DyNetVis [16]. É um software específico para visualização de redes dinâmicas que possui um paradigma diferente das demais ferramentas mencionadas. O DyNetVis utiliza o conceito de layout temporal (ao invés do estrutural) para apresentar a rede.

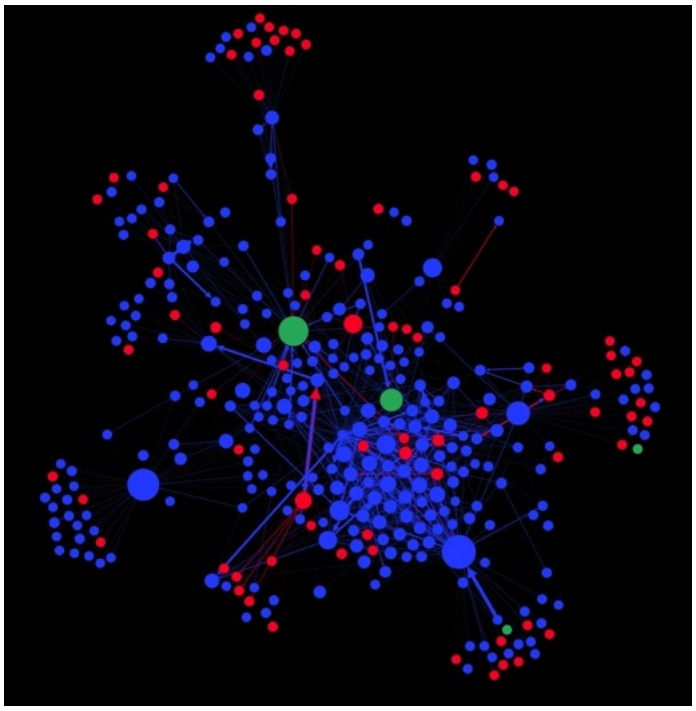

Figura 1.8. Exemplo de rede social obtida através do Gephi.

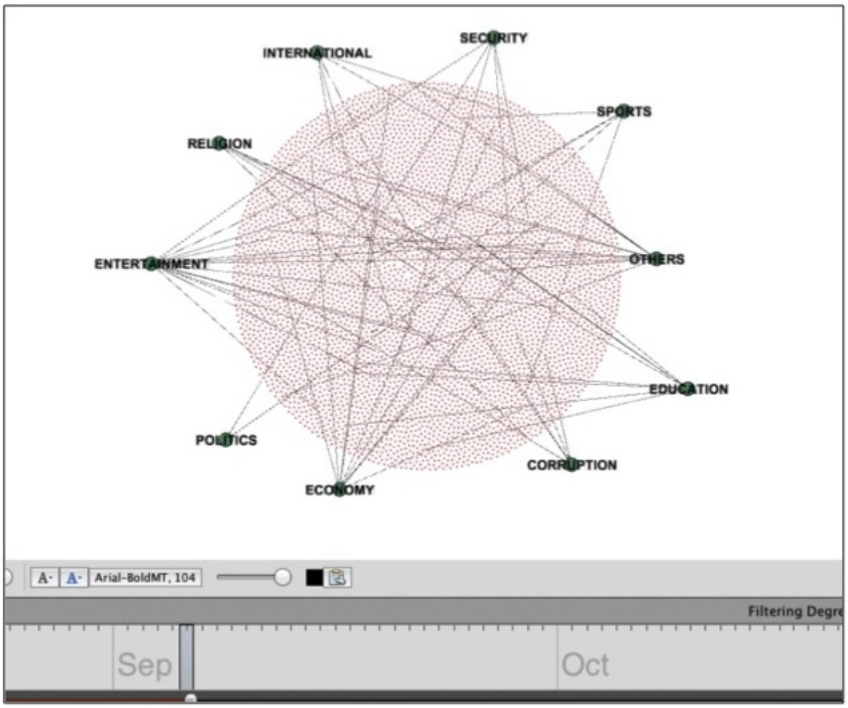

Figura 1.9. Exemplo de rede social temporal sendo analisada no Gephi.

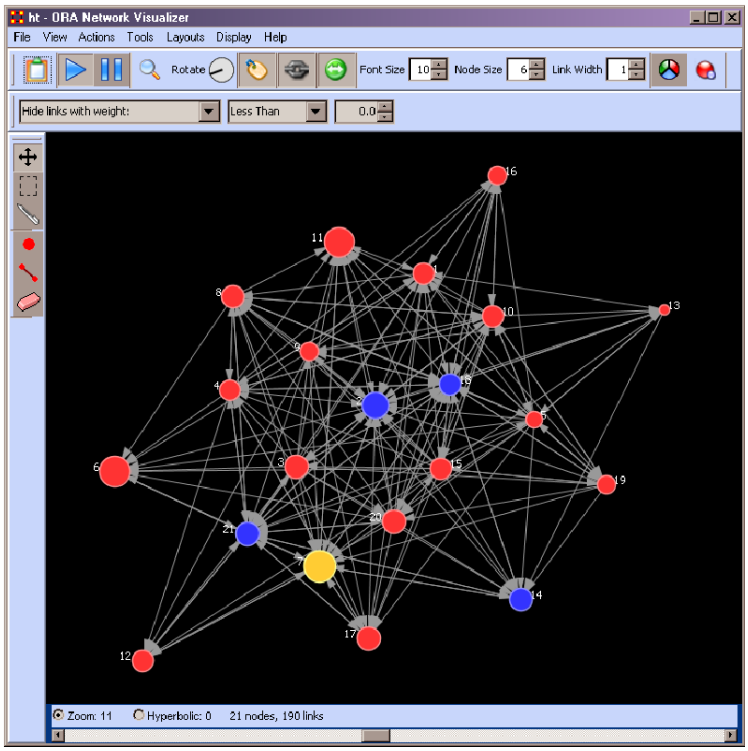

Figura 1.10. Interface da ferramenta ORA [13].

Com isso, é possível por exemplo na etapa da análise exploratória, perceber a evolução da comunicação entre nós vizinhos recorrentes (comunidades). A Figura 1.11 ilustra uma rede temporal visualizada pelo DyNetVis.

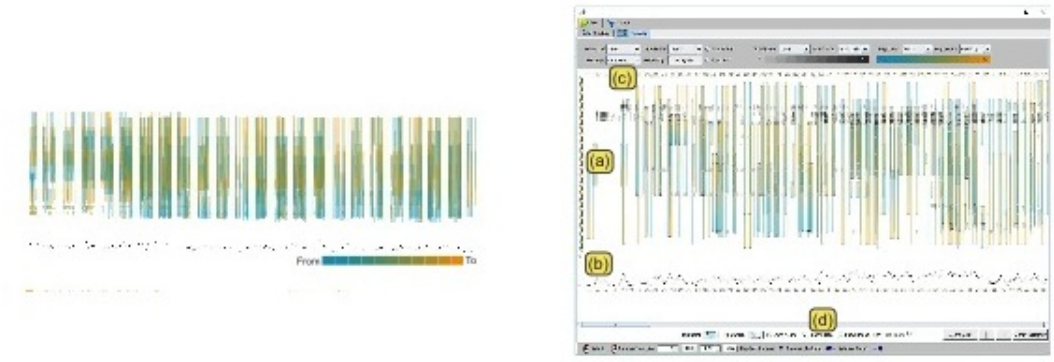

Figura 1.11. Layout temporal produzido pelo DyNetVis [16]

Outras ferramentas comuns para análise de redes sociais são R (pacotes *igraph*, *network*, *sna*), Neo4J e SNAP. Porém, quando o interesse é em análise temporal, essas ferramentas ainda não possuem flexibilidade e funcionalidades superiores àquelas acima descritas.

1.6. Estudos de Caso

Tendo como base os conceitos, estratégias e processos de análise discutidos neste capítulo, são apresentados três estudos de caso com bases de dados reais. O foco do estudo é destacar os processos utilizados na análise de cada rede para obtenção de conhecimento.

1.6.1. Rede Social do Twitter

Considerando o Twitter como domínio, nesse estudo de caso o objetivo é analisar a evolução de interações entre usuários em relação a notícias da Folha de São Paulo. Parte desse estudo foi conduzida no trabalho [21]. Na rede social, os nós são usuários do Twitter e uma aresta dirigida de u para v representa que v retweetou u[2]. A Tabela 1.1 sumariza as estatísticas da rede em questão.

Tabela 1.1. Estatísticas da rede social temporal do Twitter.

Conteúdo da Rede	
Domínio	Notícias da Folha de S. Paulo no Twitter
Política de crawling	tweets com menção @folha
Intervalo de tempo	8/8/2016 - 9/11/2016
# total tweets	1,771,435
# tweets *retweetados*	150,822
Topologia da rede	
# nós	292,310
# arestas temporais (retweets)	1,392,841

O conteúdo de notícias da rede está estruturado através de tópicos que sumarizam o assunto de um determinado tweet. Por exemplo, segurança, corrupção, política, esportes etc. A Figura 1.12 ilustra a evolução de uma amostra da rede. Note que a análise da evolução torna-se mais rica com a inserção das informações sobre os tópicos na rede (mais especificamente, nas arestas).

Considerando a metodologia para entendimento do problema e descoberta de conhecimento em redes (Seção 1.1.1), o próximo passo de análise dessa rede é explorar métricas de centralidade, tanto temporais quanto estáticas. As métricas obtidas estão sumarizadas na Tabela 1.2. Em geral os nós dessa rede possuem baixas centralidades, principalmente closeness, pois os retweets estão sempre vinculados ao usuário que originalmente postou. Não é uma cascata de interações.

Tabela 1.2. Medidas de centralidade da rede social temporal do Twitter.

Métricas da rede	estática	temporal
Média do tamanho do caminho mais curto	12.31	5 dias
Média do grau	4.76	-
Média do closeness	1.01	2.44
Média do betweenness	0.0056	0.0233

[2]Retweet é um compartilhamento que um usuário u faz de um tweet originalmente postado por outro usuário v

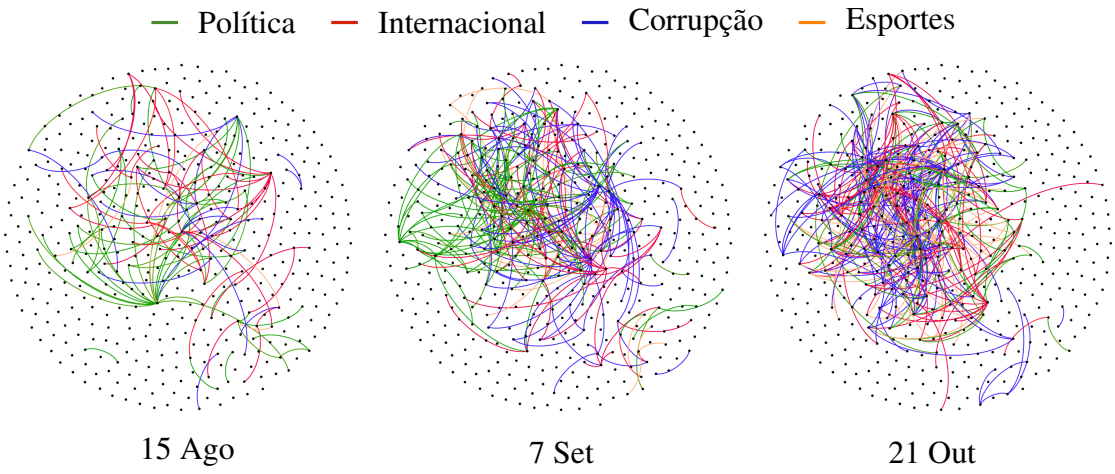

Figura 1.12. *Snapshots* de amostras da rede de retweets. Nós são usuários do Twitter. Uma ligação de u_1 para u_2 significa que u_2 retweetou no instante t algum texto originalmente postado por u_1. As cores representam tópicos que usuários estão falando sobre no instante t. As amostras foram obtidas filtrando os nós com grau entre 50-22000 e as arestas representando os 4 assuntos mais populares. Cada snapshot corresponde a 1 dia de granularidade de tempo.

Tendo então, uma ideia da semântica da rede, uma visualização de amostras dessa rede e algumas estatísticas de métricas de centralidade como um todo, é possível extrair diversas informações. Alguns dos principais conhecimentos são listados a seguir:

- A centralidade closeness está relacionada com a visibilidade de um nó na rede. É a capacidade que um nó tem de atingir os demais de maneira rápida. Portanto, um alto closeness significa uma boa capacidade de espalhar informações. Nesse contexto, é possível então identificar 3 tipos de usuários: consumidores, produtores e consumidores&produtores. Consumidores são aqueles que na maioria das vezes apenas retweetam, não publicando nenhum novo conteúdo. Geralmente, têm um baixo closeness. Produtores são aqueles que sempre estão publicando novos conteúdos, com closeness médio. Por fim, os consumidores&produtores têm um alto grau de atividade na rede, tweetando e retweetando o tempo todo. Esse tipo de usuário possui os mais altos valores de closeness.

- A predominância dos assuntos mais comentados na rede varia a cada dia. Tal efeito é intuitivo, já que reflete o dia-a-dia do contexto de notícias: a cada nova manchete, novos comentários e interações relacionados a ela surgem, trazendo um caráter natural de evolução dos assuntos predominantes na rede.

- É possível identificar quais os tópicos favoritos de cada usuário. Se um determinado usuário sempre interage na rede com o mesmo assunto, já foi possível extrair um conhecimento personalizado em relação àquele usuário.

1.6.2. Rede Social de CDRs

Call Detail Records (CDRs) são registros de ligações telefônicas em empresas de telecomunicações. Em geral, as principais informações contidas em um CDR são: o número

que originou a chamada (número de A), o número que recebeu a chamada (número de B), o instante que a chamada ocorreu e a duração. Uma rede social de CDRs é aquela em que os nós são os números de telefone e as arestas são dirigidas indicando que um número ligou para o outro. Esse estudo de caso foi extraído do trabalho [25]. A Tabela 1.3 sumariza as estatísticas da rede em questão.

Tabela 1.3. Estatísticas da rede social de CDRs.

Descrição da Rede	
Domínio	Ligações telefônicas de uma empresa de telecom
Intervalo de tempo	31 dias
Granularidade do processamento	1 dia
Característica do processamento	*streaming* (fluxo contínuo)
Topologia da rede	
# ligações	386,492,749
# números de telefone	11,916,442

Nesse estudo, o foco está no volume de dados sendo processado e na escolha da abordagem de processamento em *streams* para extrair conhecimento temporal da rede. As Figuras 1.13(a) e 1.13(b) descrevem algumas das características principais da rede em termos de evolução.

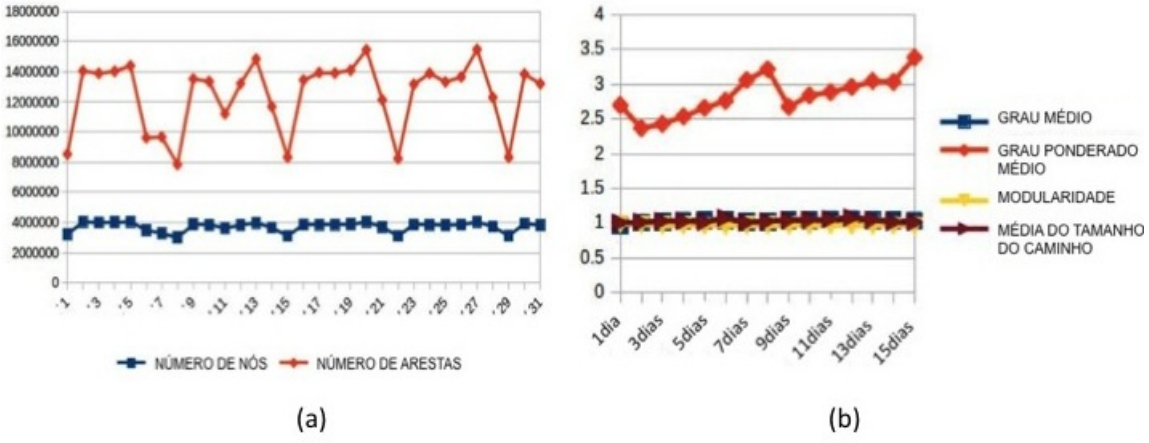

Figura 1.13. Traduzida de [25]. Evolução da rede de CDRs. (a) Evolução dos nós e arestas ao longo de 31 dias. (b) Evolução de métricas de centralidade e modularidade globais da rede.

Quanto à interpretação, é possível enumerar o seguinte conhecimento extraído:

- Existe um padrão de ligações telefônicas ao longo do mês. A cada intervalo de 7 dias o número de ligações diminui indicando os dias de domingo.
- As métricas de rede testadas se mantêm constantes ao longo do período analisado, indicando que a operadora de telefonia conseguiu manter naquele mês os seus clientes e nada de diferente ocorreu para fazer tais clientes mudarem seus padrões de comunicação.

Nesse cenário, então, além da visualização da rede (não ilustrada aqui por restrições de espaço) foi utilizada a estratégia de coletar a cada dia estatísticas da rede para entender sua evolução estrutural. Logo, é uma análise puramente estrutural, diferente da rede do Twitter em que o conteúdo da rede (tópicos) foram utilizados para extração do conhecimento. Perceba que uma entrega de um cientista de dados em rede para a empresa deve ter sempre como foco análises que possam gerar valor ao negócio de telecom.

1.6.3. Rede Social de Músicas

Esse estudo de caso descreve os resultados obtidos no trabalho [19]. A rede social nesse cenário foi obtida a partir de dados da plataforma de músicas Last.fm. Os nós são usuários da plataforma e as arestas são relações de seguidores/seguidos assim como no Twitter. O contexto é então, uma rede de amigos orientada por gostos musicais, onde cada aresta possui a indicação do instante em que a amizade foi criada. As estatísticas estão na Tabela 1.4.

Tabela 1.4. Estatísticas da rede social de músicas.

Descrição da Rede	
Domínio	Relações de amizade na plataforma de músicas Last.fm
Intervalo de tempo	1/1/2002 - 31/12/2011
Granularidade do processamento	1 dia
Topologia da rede	
# usuários	71,000
# arestas	285,241

Utilizando a mesma estratégia de análise descrita no caso de uso da rede de CDRs, uma primeira análise exploratória pode ser descrita com gráficos que mostram a evolução do comportamento da rede. As Figuras 1.14(a) e 1.14(b) descrevem o comportamento da rede em questão.

Assim como no contexto da rede do Twitter, essa análise foi feita acrescentando aos nós (usuários) informações sobre seus comportamentos – quais as músicas, playlists e artistas foram ouvidos. O objetivo desse estudo, retirado do trabalho [19], foi validar o nível de influência que uma rede de amizades pode exercer sobre os gostos musicais. Retomando à discussão proposta sobre a sequência básica de passos para análise de redes sociais temporais, nesse estudo de caso foram observadas as estatísticas, evolução do comportamento da rede e métricas. Só depois de entender bem a rede que se tem em mãos é que análises mais profundas e específicas – como o grau de influência dos amigos nas escolhas musicais – devem começar.

1.7. Considerações Finais

Neste capítulo foram discutidas estratégias para extração de conhecimento em redes sociais temporais. Tais redes representam a evolução das interações entre entidades ao longo do tempo, podendo ser aplicadas em contextos como Twitter, redes de ligações telefônicas e redes de gostos musicais. Primeiro foram definidos diversos conceitos por trás da ideia temporal de análise de redes. Tais conceitos mostram que as métricas de centralidade

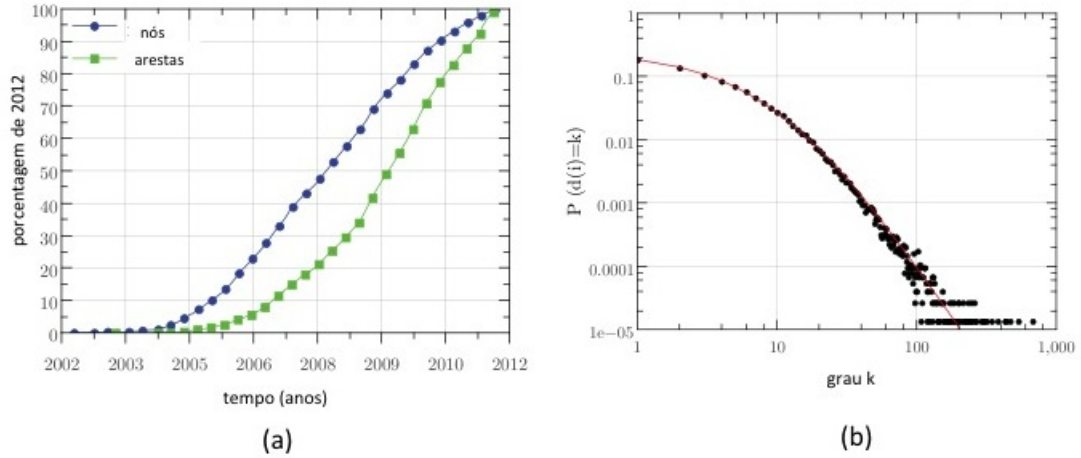

Figura 1.14. Traduzida de [19]. Evolução da rede social de amizades da plataforma Last.fm. (a) Evolução dos nós e arestas ao longo dos anos. (b) Distribuição dos graus dos nós da rede.

baseadas em caminhos devem ser revisitadas e repensadas no contexto temporal. Em seguida, foram discutidas estratégias para processamento de redes que evoluem à medida que o tempo passa, variando entre estratégias para redes que evoluem lentamente e para redes com atualização constante, em fluxos de arestas. Uma vez apresentados os conceitos, em caráter prático, foram mostradas técnicas de visualização que podem fornecer *insights* no momento da análise da rede, bem como ferramentas que auxiliam tais atividades. Por fim, os conceitos, técnicas e ferramentas foram aplicadas em três estudos de caso que ilustram como deve ser uma análise para descoberta de conhecimento em redes sociais temporais.

Referências

[1] Adedoyin-Olowe, M., Gaber, M.M., Stahl, F.: A Survey of Data Mining Techniques for Social Media Analysis. Journal of Data Mining & Digital Humanities 2014 (Jun 2014), http://jdmdh.episciences.org/18

[2] Ahmed, N.K., Neville, J., Kompella, R.: Network sampling: From static to streaming graphs. ACM Trans. Knowl. Discov. Data 8(2), 7:1–7:56 (Jun 2013), http://doi.acm.org/10.1145/2601438

[3] Bach, B., Pietriga, E., Fekete, J.D.: Visualizing dynamic networks with matrix cubes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 877–886. CHI '14, ACM, New York, NY, USA (2014)

[4] Bastian, M., Heymann, S., Jacomy, M.: Gephi: An open source software for exploring and manipulating networks (2009)

[5] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Physics Reports 424(4–5), 175–308 (2006)

[6] Carley, K.M., Pfeffer, J.: Dynamic network analysis (dna) and ora. Advances in Design for Cross-Cultural Activities Part I p. 265Ð274 (2012)

[7] Costa, L.F., Rodrigues, F., Travieso, G., Villas Boas, P.R.: Characterization of complex networks: a survey of measurements. Advances in Physics 56(1), 167–242 (2007)

[8] Franca, T.C., de Faria, F.F., Rangel, F.M., de Farias, C.M., Oliveira, J.: Big social data: Principios sobre coleta, tratamento e analise de dados sociais. In: Anais do SBBD. pp. 1–40 (2014)

[9] Guha, S., McGregor, A.: Graph synopses, sketches, and streams: A survey. Proc. VLDB Endow. 5(12), 2030–2031 (Aug 2012), http://dx.doi.org/10.14778/2367502.2367570

[10] Holme, P.: Analyzing temporal networks in social media. Proceedings of the IEEE 102(12), 1922–1933 (2014)

[11] Holme, P., Saramaki, J.: Temporal networks. Physics Reports 519(3), 97–125 (2012)

[12] Huhtamäki, J.: Visualizing co-authorship networks for actionable insights: Action design research experiment. In: Proceedings of the 20th International Academic Mindtrek Conference. pp. 208–215. AcademicMindtrek '16, ACM, New York, NY, USA (2016)

[13] Huisman, M., Duijn, M.A.J.V.: A Reader's Guide to SNA Software. Sage (2011)

[14] Kempe, D., Kleinberg, J., Kumar, A.: Connectivity and inference problems for temporal networks. In: Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing. pp. 504–513. ACM (2000)

[15] Kim, H., Anderson, R.: Temporal node centrality in complex networks. Phys. Rev. E 85, 026107 (Feb 2012)

[16] Linhares, C.D.G., Travençolo, B.A.N., Paiva, J.G.S., Rocha, L.E.C.: Dynetvis: A system for visualization of dynamic networks. In: Proceedings of the Symposium on Applied Computing. pp. 187–194. SAC '17, ACM, New York, NY, USA (2017)

[17] Nicosia, V., Tang, J., Mascolo, C., Musolesi, M., Russo, G., Latora, V.: Temporal Networks, chap. Graph Metrics for Temporal Networks, pp. 15–40. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)

[18] Oezbek, C., Prechelt, L., Thiel, F.: The onion has cancer: Some social network analysis visualizations of open source project communication. In: Proceedings of the 3rd International Workshop on Emerging Trends in Free/Libre/Open Source Software Research and Development. pp. 5–10. FLOSS '10, ACM, New York, NY, USA (2010)

[19] Pálovics, R., Benczúr, A.A.: Temporal influence over the last.fm social network. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 486–493. ASONAM '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2492517.2492532

[20] Pan, R.K., Saramäki, J.: Path lengths, correlations, and centrality in temporal networks. Phys. Rev. E 84 (2011)

[21] Pereira, F.S.F., de Amo, S., Gama, J.: Detecting events in evolving social networks through node centrality analysis. Workshop on Large-scale Learning from Data Streams in Evolving Environments co-located with ECML/PKDD (2016)

[22] Pereira, F.S.F., Amo, S., Gama, J.: Evolving centralities in temporal graphs: a twitter network analysis. In: Mobile Data Management (MDM), 2016 17th IEEE International Conference on (2016)

[23] Rossi, R., Gallagher, B., Neville, J., Henderson, K.: Role-dynamics: Fast mining of large dynamic networks. In: Proceedings of the 21st International Conference on World Wide Web. pp. 997–1006. WWW'12 Companion, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2187980.2188234

[24] Srivastava, J.: Data mining for social network analysis. In: 2008 IEEE International Conference on Intelligence and Security Informatics. pp. 23–24. IEEE, Taiwan (June 2008)

[25] Tabassum, S., Gama, J.: Sampling massive streaming call graphs. In: Proceedings of the 2016 ACM Symposium on Applied Computing. pp. 923–928. SAC '16, ACM, New York, NY, USA (2016)

[26] Tang, J., Musolesi, M., Mascolo, C., Latora, V.: Temporal distance metrics for social network analysis. In: Proceedings of the 2nd ACM Workshop on Online Social Networks. pp. 31–36. WOSN '09 (2009)

[27] Viswanath, B., Mislove, A., Cha, M., Gummadi, K.P.: On the evolution of user interaction in facebook. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09) (August 2009)

[28] Wei, W., Carley, K.M.: Measuring temporal patterns in dynamic social networks. ACM Trans. Knowl. Discov. Data 10(1), 9:1–9:27 (Jul 2015)

[29] Wu, H., Cheng, J., Huang, S., Ke, Y., Lu, Y., Xu, Y.: Path problems in temporal graphs. Proceedings of the VLDB Endowment 7(9), 721–732 (2014)

[30] Zafarani, R., Abbasi, M.A., Liu, H.: Social Media Mining: An Introduction. Cambridge University Press, New York, NY, USA (2014)

1.8. Sobre os Autores

Fabíola Souza Fernandes Pereira. Doutoranda em Ciência da Computação na Universidade Federal de Uberlândia (UFU), com período sanduíche no LIAAD, um grupo pertencente ao INESC TEC, Portugal. Possui graduação (2009) e mestrado (2011) em Ciência da Computação também pela UFU. É autora de artigos peer-reviewed nas áreas de redes temporais, análise de redes sociais e preferências do usuário. Atuou como chair da special session em redes evolutivas (EvoNets) na conferência DSAA'17.

João Gama. É professor associado da Faculdade de Economia, Universidade do Porto. É pesquisador e vice-diretor do LIAAD, um grupo pertencente ao INESC TEC. Obteve o título de Ph.D. pela Universidade do Porto em 2000. Tem trabalhado em diversos projetos nacionais e europeus em sistemas de aprendizado incremental e adaptativo, descoberta de conhecimento ubíquo, aprendizado a partir de dados massivos e em fluxo, etc. É autor de diversos livros em Mineração de Dados e de mais de 250 artigos peer-reviewed nas áreas de aprendizado de máquina, mineração de dados e data streams.

Gina Maira Barbosa de Oliveira. Bolsista de produtividade do CNPq de 2001 a 2017 (PQ-2). Possui graduação em Engenharia Elétrica pela Universidade Federal de Uberlândia (1990), mestrado em Engenharia Eletrônica e Computação pelo Instituto Tecnológico de Aeronáutica (1992) e doutorado em Engenharia Eletrônica e Computação pelo Instituto Tecnológico de Aeronáutica (1999). Pós-doutorado de 07/2013 a 07/2014 na Heriot-Watt University (Edinburgh-Scotland) na área de robótica bio-inspirada. Atualmente é professora associada da Universidade Federal de Uberlândia. Tem experiência na área de Ciência da Computação, atuando principalmente nos seguintes temas: algoritmos genéticos, autômatos celulares, computação evolutiva, computação bio-inspirada, robótica bio-inspirada e inteligência artificial.