Potencial de influência em publicações usando território causal

Diogo F. S. Ramos[1] **and Jesús P. Mena-Chalco**[1]

[1]Centro de Matemática, Computação e Cognição
Universidade Federal do ABC (UFABC) – Santo André, SP – Brasil

diogo.ramos@ufabc.edu.br, jesus.mena@ufabc.edu.br

Abstract. *Temporal graphs are an extension of conventional graphs to represent time related events. In this article, we introduce an algorithm to compute a member set from a network which can be directly or indirectly influenced with a computational cost $O(VC)$, as found in the literature, but with a smaller average. We formally define the* Causal Territory *concept. We made experiments on the first five years of researchers' careers present on DBLP. The main results show that: (i) the size of causal territories are increasing; (ii) the concentration of potentially influenced researchers are stabilizing; and (iii) researchers are increasing their influence reach.*

Resumo. *Grafos temporais são uma extensão dos grafos convencionais para representar eventos que ocorrem no tempo. Neste trabalho introduzimos um algoritmo que calcula o conjunto de membros de uma rede que são influenciados direta ou indiretamente com custo computacional $O(VC)$, tal qual encontrado na literatura, mas com média menor. Formalmente definimos o conceito de* Território Causal. *Realizamos experimentos sobre os primeiros cinco anos de carreira dos pesquisadores presentes na DBLP. Os principais resultados indicam que: (i) o tamanho dos territórios causais vem aumentando; (ii) a concentração da quantidade de pesquisadores potencialmente influenciados vem se estabilizando; e (iii) os pesquisadores tem aumentado o seu alcance de influência.*

1. Introdução

A análise de eventos sociais é um tema de grande interesse em diversas áreas do conhecimento, como sociologia, psicologia, antropologia e ciência política. Entender como as pessoas interagem em eventos como reuniões, festas e conferências pode fornecer percepções valiosas sobre a dinâmica social e as relações interpessoais [Barros et al. 2021]. No entanto, a análise de eventos sociais pode ser desafiadora devido à sua natureza complexa e dinâmica. Muitas vezes, os participantes mudam ao longo do tempo [Palla et al. 2007] e as interações entre eles podem ser difíceis de quantificar ou visualizar [Greene et al. 2010].

Nesse contexto, a teoria dos grafos temporais surge como uma ferramenta poderosa para modelar e analisar eventos sociais. Grafos temporais são uma extensão dos grafos convencionais que permitem representar eventos que ocorrem ao longo do tempo [Holme and Saramäki 2012]. Eles podem ser utilizados para visualizar as relações entre os participantes em um evento social e para identificar padrões interessantes na dinâmica da rede [Holme 2014]. Tal qual grafos estáticos (teoria clássica de grafos), os grafos temporais possuem métodos para calcular as recíprocas de métricas como centralidade de grau ou de intermediação [Tang et al. 2010].

Uma característica de eventos sociais é que os seus participantes são capazes de influenciar uns ao outros [Yu et al. 2015]. Essa influência pode ser direta ou indireta. No caso de influência direta, um participante interage diretamente com um outro. No caso de influência indireta, um dos participantes carrega consigo a influência que teve de uma primeira interação com um participante para uma nova interação com um novo participante. Mesmo que o primeiro participante não interaja diretamente com o terceiro, ele tem o *potencial* de influenciá-lo.

Neste trabalho, apresentamos um algoritmo capaz de calcular o conjunto de todos os membros de uma rede que possam ser direta ou indiretamente influenciados por um membro qualquer da rede. Chamamos de *território causal* o conjunto de todos os membros que possam ter sido influenciados por um membro arbitrário. O algoritmo também calcula o momento mais cedo no qual um membro possa ter sido influenciado. O cálculo de caminhos (chamados de *jornadas*) em grafos temporais já foi abordado por outros pesquisadores [Wu et al. 2014]. A vantagem de nossa abordagem é a que não é necessário explorar toda a rede e que todos os membros potencialmente influenciados são incluídos na resposta.

Por usar um grafo temporal, a nossa técnica permite identicar os vértices que potenticalmente foram influenciados e, não menos importante, por exclusão identificar os vértices que não foram influenciados. Em um grafo estático, se há um caminho de um vértice u para um vértice v, há também um caminho de v para u. O mesmo não ocorre em grafos temporais. Por exemplo, suponha um encontro entre três pessoas, p_1, p_2 e p_3, em sequência, ou seja, p_1 encontrou p_2 antes de p_2 encontrar p_3. Suponha também que p_3 está doente. Ao modelarmos o problema usando um grafo temporal, conseguiremos concluir que é impossível p_3 infectar p_1, pois a interação entre p_2 e p_3 ocorreu depois da interação entre p_1 e p_2.

Com o algoritmo de cálculo do território causal, exploramos a base de dados "Digital Bibliography & Library Project" - DBLP [Ley 2002] e caracterizamos a evolução dos territórios causais de jovens pesquisadores a cada ano. Finalmente, é importante destacar que a justificativa para este trabalho é que ele oferece uma perspectiva aprimorada sobre a análise de eventos sociais, permitindo uma compreensão mais profunda da dinâmica das redes sociais e das relações interpessoais.

2. Grafo temporal

Um *grafo temporal* é um grafo que considera a ordem temporal dos eventos [Michail 2016]. Ele possui *vértices* e *contatos*. Os contatos são eventos entre vértices que ocorrem em determinados *momentos*. Um grafo temporal pode ser especificado por um conjunto de vértices e contatos, $\{V, C\}$, no qual V representa o conjunto de vértices $\{u_1, u_2, u_3, \ldots, u_n\}$ e C o conjunto de contatos $\{c_1, c_2, c_3, \ldots, c_m\}$, sendo cada contato c_i uma tripla (u, v, m), na qual u e v são vértices e m é o momento no qual o contato ocorreu. Os contatos podem ser dirigidos ou não.

Uma sequência de contatos entre dois vértices dentro de um grafo temporal é chamado de *jornada*. Os momentos de contato de uma jornada precisam ser estritamente crescentes. Por exemplo, na Figura 1 temos um grafo temporal com dois vértices, u e v, e dois contatos, $(u, v, 1)$ e $(v, w, 2)$. Há uma jornada de u a w definida pela sequência de contatos $\langle(u, v, 1), (v, w, 2)\rangle$; porém, não há uma jornada de w a u, pois o contato entre w

e v ocorreu após o contato entre v e u.

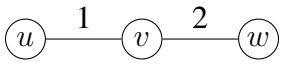
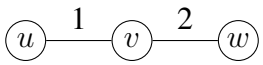


Figura 1. Exemplo de grafo temporal. O grafo possui três vértices $\{u, v, w\}$ e dois contatos $\{(u, v, 1), (v, w, 2)\}$.

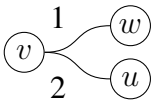
3. Território causal

Um território causal é o conjunto de vértices que pode ser influenciado direta ou indiretamente por um vértice arbitrário. Mais especificamente, é uma árvore enraizada com todos os vértices possíveis de serem alcançados por jornadas com os menores momentos possíveis de contato.

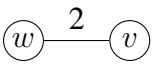
Considere o exemplo de grafo temporal da Figura 1. Nele, há três vértices e dois contatos. Os territórios causais dos vértice u e v possuem três vértices (u, v e w) enquanto o território causal do vértice w possui dois vértices (Figura 2). O território causal do vértice w possui um vértice a menos do que os demais porque, quando o contato entre ele e v ocorre (momento 2), já é tarde demais para que ocorra o contato entre v e u (momento 1) e assim w não tem influência sobre o vértice u.



(a) Território causal do vértice u



(b) Território causal do vértice v



(c) Território causal do vértice w

Figura 2. Territórios causais dos vértices do exemplo do Figura 1.

4. Algoritmo do território causal

O território causal pode ser calculado de diferentes maneiras. Nesta seção descrevemos o algoritmo que foi projetado baseado no algoritmo de Dijkstra [Dijkstra 1959].

A função *CAUSAL-TERRITORY* (Algoritmo 1) implementa o cálculo do território causal. Como entrada, a função recebe um grafo G, um vértice s e um período dentro do qual o território causal deve estar contido, representado pelos parâmetros t_{start} e t_{end}. Essa função usa três variáveis auxiliares:

π Tabela relacionando vértices com seus pais;
q Fila de prioridade que ordena os vértices por ordem de momento de descobrimento;
t Tabela determinando se um vértice já foi explorado.

Inicialmente o vértice inicial s é adicionado à fila q com momento $t_{start} - 1$, pois assim somente contatos com momentos posteriores a esse podem ser adicionados ao território causal.

Entra-se então em um laço do qual só se sai quando todos os vértices do território causal forem encontrados. Nesse laço os vértices com os momentos menores são explorados primeiros. A cada volta, os vértices adjacentes são descobertos e, caso não tenham sido visitados ainda, os seus momentos de descoberta são registrados de acordo

```
function CAUSAL-TERRITORY(G, s, t_start, t_end):
    π, q, t;
    ENQUEUE(q, s, t_start − 1);
    until EMPY(q) do
        u, m_u ← DEQUEUE(q);
        forall v ∈ ADJ(G, u) do
            if v ∉ t then
                RELAX(u, m_u, v);
        t[u] ← m_u;
    return π
```

Algoritmo 1: Território causal. Esta é a função de entrada do algoritmo que calcula o território causal de um vértice u dentro de um grafo temporal G. Os parâmetros t_{start} e t_{end} permitem estabelecer um período no qual o território causal pode ser explorado ($t_{start} \leq m < t_{end}$).

```
procedure RELAX(u, m_u, v):
    m_uv ← {m ∈ MOMENTS(G, u, v) : m_u < m < end};
    if m_uv ≠ ∅ then
        m_min ← MIN(m_uv);
        if v ∈ q then
            if m_min < q_v then
                π[v] ← u;
                q[v] ← m_min;
        else
            π[v] ← u;
            ENQUEUE(q, v, m_min);
```

Algoritmo 2: Procedimento auxiliar *RELAX*. Neste procedimento, os vértices adjacentes passados pela função *CAUSAL-TERRITORY* são processados. Eles podem ser: ignorados; ter seus momentos de descoberta alterados caso estejam na fila q; ser adicionados à fila q.

com o procedimento *RELAX* (Algoritmo 2). Após todos os vértices adjacentes terem sido encontrados e processados, o vértice desenfileirado é marcado como visitado na tabela t. Ao final da execução, a tabela π é retornada.

O procedimento *RELAX* determina o que fazer com os vértices adjacentes encontrados dentro do laço da função *CAUSAL-TERRITORY*. Primeiramente os momentos possíveis entre os vértices u e v são filtrados. Caso não haja nenhum, nada é feito. Caso haja, escolhe-se o menor momento possível. Caso o vértice adjacente v esteja dentro da fila para ser posteriormente processado, verifica-se se não é possível trocá-lo de posição, isto é, se o momento encontrado não é menor que o momento que o vértice possui dentro da fila. Em caso positivo, troca-se seu pai e seu momento na fila. Em caso negativo, nada se faz. É importante destacar que, caso o vértice não esteja na fila, define-se o vértice u como seu pai e enfileira-se o vértice v.

O território causal é calculado a partir do vértice de interesse e só é necessário

explorar os vértices que estão na vizinhança do território causal. A complexidade do algoritmo é a mesma da encontrada na literatura ($O(VC)$); porém, por ser necessário somente explorar os vértices que estão na vizinhança do território causal final, o seu tempo de execução médio é menor. Desse modo, quanto mais esparso o grafo menor será o tempo de execução.

5. Grafo temporal de coautoria científica

As interações acadêmicas entre pesquisadores podem ser desenvolvidas em diferentes formas, assim como em graus de intensidade que perduram ao longo do tempo [Katz and Martin 1997]. A interação acadêmica que é comumente mensurada na análise de redes sociais acadêmicas [Kong et al. 2019] se baseia nas publicações científicas feitas com participação de diferentes coautores.

No contexto da colaboração científica, um grafo temporal de coautoria é uma modelagem de uma rede de coautoria científica usando grafo temporal não dirigido. Nesse grafo temporal, os vértices representam os coautores de uma publicação científica e os contatos representam o momento no qual a publicação foi publicada. Os momentos de contato são os anos da publicação. Um momento mais preciso seria desejável; porém, nem toda publicação indica o mês ou dia de sua publicação (nas bases de dados bibliométricas, como DBLP, *Scopus* ou *Web of Science*, a granularidade de registro das produções científica se refere ao ano de publicação).

A Figura 3 apresenta um exemplo ficcional de rede de coautoria. Nela, há quatro artigos publicados entre os momentos 1 e 4 e seis autores. A definição formal do grafo temporal é $G = (V, C)$ no qual:

$$
\begin{aligned}
V =& \{a_1, a_2, a_3, a_4, a_5, a_6\} \\
C =& \{(a_1, a_2, 1), (a_2, a_4, 2), (a_1, a_2, 3), (a_1, a_3, 3), \\
& (a_2, a_3, 3), (a_4, a_5, 4), (a_4, a_6, 4), (a_5, a_6, 4)\}
\end{aligned}
$$

Artigo	Momento	Autores
A_1	1	a_1, a_2
A_2	2	a_2, a_4
A_3	3	a_1, a_2, a_3
A_4	4	a_4, a_5, a_6

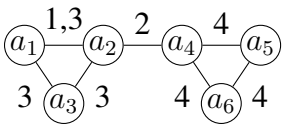
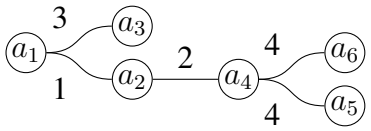
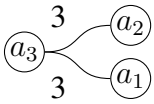


Figura 3. Exemplo de coautoria em quatro artigos modelados usando grafo temporal. Cada vértice representa um autor; os contatos, as coautorias; os momentos dos contatos, o ano de publicação do artigo.

Sendo um grafo temporal, é possível calcularmos os territórios causais de seus vértices. Como exemplo, considere os vértices a_1 e a_3: o território causal de a_1 ($V = \{a_1, a_2, a_3, a_4, a_5, a_6\}$ e $C = \{(a_1, a_2, 1), (a_1, a_3, 3), (a_2, a_4, 2), (a_4, a_5, 4), (a_4, a_6, 4))\}$) possui todos os vértices do grafo enquanto o território causal de a_3 ($V = \{a_3, a_1, a_2\}$ e $C = \{(a_3, a_1, 3), (a_3, a_2, 3)\}$) possui somente três vértices (Figura 4). Note que o vértice a_1 possui um território com altura três enquanto o vértice a_3 possui um território com altura um.



(a) Território causal do vértice a_1.



(b) Território causal do vértice a_3.

Figura 4. Diferença entre territórios causais da mesma rede de coautoria. Os vértices a_1 e a_3 possuem territórios causais diferentes mesmo sendo da mesmo rede de coautoria (grafo da Figura 3) e estarem cada um dentro do território causal do outro.

6. Método

Para o desenvolvimento[1] de todos os experimentos deste trabalho usamos a base de dados de publicações da ciência da computação "Digital Bibliography & Library Project" (DBLP, https://dblp.org/xml/release/). Usamos a versão da base do dia 27 de março de 2023, a qual possui:

- 6 528 720 Publicações científicas;
  - 3 215 335 artigos em congresso;
  - 3 117 167 artigos em revistas;
  - 196 218 dissertações, teses etc;
- 3 215 112 pesquisadores.

Para o nosso estudo, usamos somente os artigos publicados em congressos e em revistas científicas, apesar da base DBLP listar diversos tipos de publicações científicas. A DBLP também lista pesquisadores que não conseguiu individualizar, agrupando todas as publicações, bem como grupos de pesquisa, como autores individuais. Para o nosso estudo, filtramos os pesquisadores que eram de fato indivíduos (Figura 5), dado que este trabalho visa observar o comportamento individual de jovens pesquisadores (nos primeiros anos após sua primeira publicação).

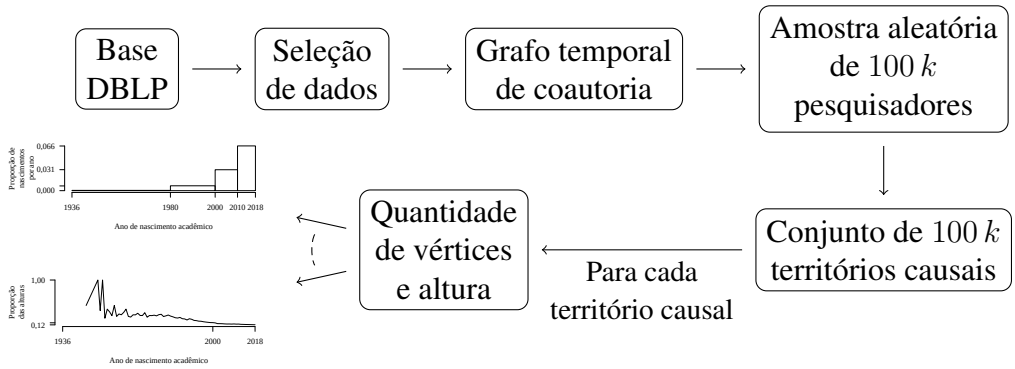


Figura 5. Delineamento de pesquisa.

A partir dos dados filtrados, calculamos o grafo temporal de coautoria. Desse grafo, retiramos uma amostra aleatória com 100 k pesquisadores nascidos academica-

[1] O algoritmo foi implementado na linguagem de programação Common Lisp e executado usando SBCL versão 2.3.4.

mente[2] até 2018. Para cada pesquisador, calculamos o seu território causal por 5 anos a partir de seu nascimento acadêmico e, de seu território causal, calculamos as métricas: quantidade de vértices e a altura. Dessas métricas, geramos gráficos para analisar suas evoluções no tempo.

O grafo temporal gerado possuía $3\,166\,303$ vértices e $25\,871\,420$ contatos. O primeiro contato possuía o momento 1938 e o último contato possuía o momento 2023. O vértice com a maior quantidade de contatos possuía 2099 contatos.

Por fim, plotamos os gráficos das duas métricas por tempo. Calculamos tanto a evolução da média das métricas por cada ano de nascimento acadêmico como a concentração dessas métricas nos $10\,\%$ de pesquisadores com as maiores métricas de cada ano.

7. Resultados

Houve um grande crescimento na quantidade de nascimentos acadêmicos nos últimos anos (Figura 6). A partir dos anos 2000 o crescimento é acelerado e não há sinais de arrefecimento até o último ano computado (2018): somente nos últimos 8 anos, houve aproximadamente $50\,\%$ dos nascimentos acadêmicos de todo o período analisado (de 1936 a 2018).

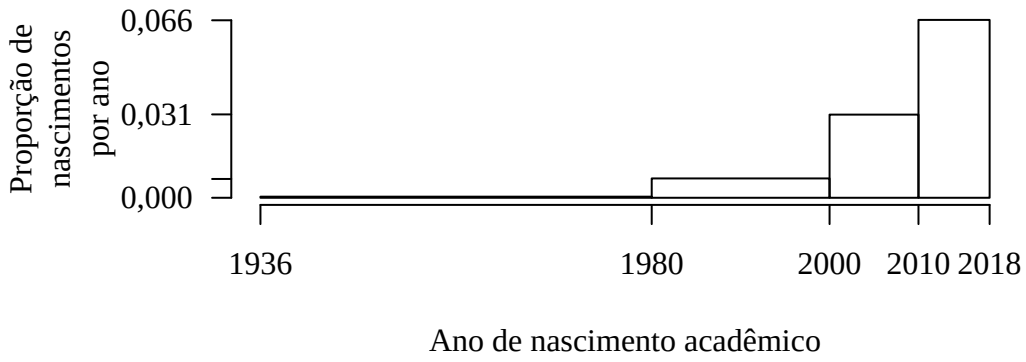


Figura 6. Evolução da quantidade de nascimentos acadêmicos por ano. No histograma, o eixo vertical representa a proporção dos nascimentos acadêmicos por ano e o eixo horizontal, os anos.

Juntamente com o aumento na quantidade de nascimentos acadêmicos, aumentou também a quantidade média da quantidade de vértices dos territórios causais dos pesquisadores por ano de nascimento acadêmico (Figura 7). Pesquisadores iniciantes passaram a poder influenciar uma quantidade cada vez maior de pesquisadores.

Apesar do aumento na quantidade nascimentos acadêmicos e na média do tamanho dos territórios causais, houve uma diminuição, mesmo que leve, na concentração do tamanho dos territórios causais nos $10\,\%$ dos pesquisadores com os maiores territórios causais nas últimas décadas (Figura 8). A partir da década de 1950, houve um aumento na concentração da quantidade de vértices dos territórios causais nos pesquisadores com os maiores territórios, mas essa tendência inverte por volta da década de 1990.

[2]A data de nascimento acadêmico de um pesquisador é a data na qual esse pesquisador publicou o seu primeiro artigo científico. Essa abordagem está de acordo com a proposta em [Nane et al. 2017].

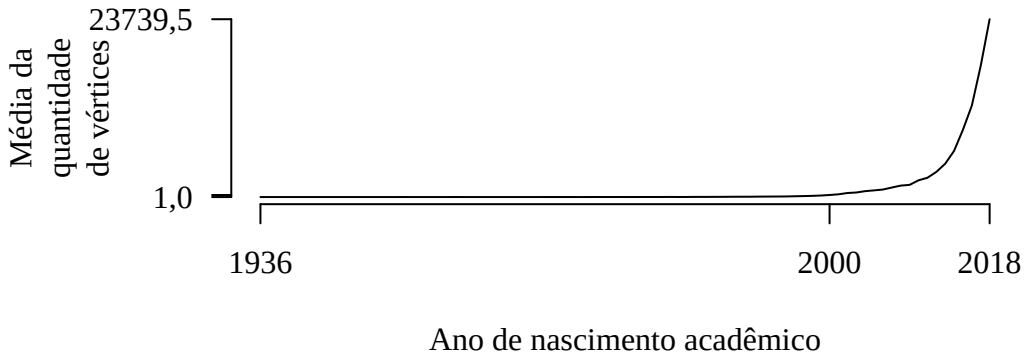


Figura 7. Evolução do potencial de influência de pesquisadores iniciantes. A quantidade média de vértices de territórios causais de pesquisadores por ano de nascimento acadêmico quantifica a quantidade média de pesquisadores que um pesquisador pode influenciar nos primeiros cinco anos após a publicação do primeiro trabalho acadêmico.

Figura 8. Concentração da quantidade de vértices nos 10 % de pesquisadores com os maiores territórios. Cada ano possui a proporção de vértices que os 10 % de pesquisadores nascidos academicamente nesse ano e com a maior quantidade de vértices em seus territórios causais possui em relação à quantidade de todos os vértices presentes em todos os territórios de pesquisadores nascidos no mesmo ano.

Um território causal pode ser visto como uma árvore, assim, tal como a quantidade de vértices em cada território causal, a altura de cada território causal também cresceu com o passar dos anos (Figura 9). Diferentemente da quantidade de vértices, não houve um crescimento acelerado nas últimas décadas, com um aumento de aproximadamente $30\,\%$ entre os anos 2000 e 2018.

Apesar do aumento na quantidade nascimentos acadêmicos e na média da altura dos territórios causais (como no caso da quantidade de vértices de cada território causal), houve uma diminuição ao longo do tempo na concentração das alturas dos territórios causais nos $10\,\%$ de pesquisadores com as maiores alturas (Figura 10).

Figura 9. Evolução do potencial de influência de pesquisadores iniciantes. Os territórios causais podem ser vistos como árvores, logo possuem altura, calculada da raiz até a folha mais longínqua. Cada ano possui a média das alturas dos territórios de todos os pesquisadores nascidos academicamente nesse ano.

Figura 10. Concentração da altura dos territórios causais nos 10 % de pesquisadores com as maiores alturas. Cada ano possui a proporção de territórios causais que os 10 % de pesquisadores nascidos academicamente nesse ano e com a maior altura de território possui em relação às alturas de todos os territórios dos pesquisadores nascidos no mesmo ano.

8. Conclusões

O nosso algoritmo para o cálculo do território causal encontra todos os vértices capazes de serem influenciados por um nó raiz no menor momento possível. A quantidade de contatos entre os vértices não é considerada, contanto que o momento da influência potencial seja o menor possível.

Usando o algoritmo e a base de dados DBLP, calculamos duas métricas dos territórios causais: quantidade de vértices e altura do território (o território causal pode ser visto como uma árvore enraizada).

A quantidade média de vértices de cada árvore aumentou com o passar dos anos. Apesar disso, a proporção da quantidade de vértices dos territórios causais dos $10\,\%$ dos pesquisadores com as maiores árvores tem se estabilizado nos últimos anos. Ainda assim, os $10\,\%$ dos pesquisadores com as maiores árvores concentram mais de $50\,\%$ de todos os vértices presentes em territórios causais. Isso significa que uma quantidade pequena de pesquisadores tem a capacidade de influenciar metade de todos os vértices.

A altura média dos territórios causais também aumentou com o passar dos anos; porém, a altura dos territórios dos $10\,\%$ de pesquisadores com os mais altos territórios concentram aproximadamente $10\,\%$ de todas as alturas. Isso deve-se ao aumento na quantidade de pesquisadores e ao limite imposto pelo nosso método: com o aumento na quantidade pesquisadores, os territórios tendem a ser mais altos e ao limitar em cinco anos os territórios causais, os territórios não podem ser mais altos do que cinco, limitando o seu tamanho total.

Finalmente, é importante destacar que as técnicas e métodos apresentados neste trabalho podem ser aplicados em uma ampla variedade de contextos. Em eventos sociais, é possível identificar quais membros possam ter sido influenciados por um outro. Em um contexto de saúde pública, é possível identificar quais indivíduos possam ter sido infectados por um outro. Em redes sociais, é possível identificar quais usuários possam ter sido influenciados por um outro. Basta que o problema seja modelado usando um grafo temporal.

Agradecimentos

Agradecemos aos pareceristas pelas críticas e sugestões ao nosso trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

Barros, C. D. T., Mendonça, M. R. F., Vieira, A. B., and Ziviani, A. (2021). A survey on embedding dynamic graphs. *ACM Comput. Surv.*, 55(1).

Dijkstra, E. W. (1959). *A Note on Two Problems in Connexion with Graphs*, page 287290. Association for Computing Machinery, New York, NY, USA, 1 edition.

Greene, D., Doyle, D., and Cunningham, P. (2010). Tracking the evolution of communities in dynamic social networks. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 176–183.

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

Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics Reports*, 519(3):97–125.

Katz, J. and Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1):1–18.

Kong, X., Shi, Y., Yu, S., Liu, J., and Xia, F. (2019). Academic social networks: Modeling, analysis, mining and applications. *Journal of Network and Computer Applications*, 132:86–103.

Ley, M. (2002). The dblp computer science bibliography: Evolution, research issues, perspectives. In Laender, A. H. F. and Oliveira, A. L., editors, *String Processing and Information Retrieval*, pages 1–10, Berlin, Heidelberg. Springer Berlin Heidelberg.

Michail, O. (2016). An introduction to temporal graphs: An algorithmic perspective. *Internet Mathematics*, 12(4):239–280.

Nane, G. F., Larivière, V., and Costas, R. (2017). Predicting the age of researchers using bibliometric data. *Journal of Informetrics*, 11(3):713–729.

Palla, G., Barabasi, A.-L., and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136):664–667.

Tang, J., Musolesi, M., Mascolo, C., Latora, V., and Nicosia, V. (2010). Analysing information flows and key mediators through temporal centrality metrics. In *3rd Workshop on Social Network Systems*.

Wu, H., Cheng, J., Huang, S., Ke, Y., Lu, Y., and Xu, Y. (2014). Path problems in temporal graphs. In *Proceedings of the Very Large Databases Endowment*, volume 7, pages 721–732. VLDB Endowment.

Yu, Z., Du, R., Guo, B., Xu, H., Gu, T., Wang, Z., and Zhang, D. (2015). Who should i invite for my party? combining user preference and influence maximization for social events. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 879883, New York, NY, USA. Association for Computing Machinery.