O nouă abordare asupra problemei comunităților

Șichet Paul-Cristian

2024

1 Abstract

Deși problema detectării comunităților este una intens studiată, având aplicații în numeroase domenii(biologie, matematică, rețele de socializare, psihologie, etc), nu are o formulare clară și concisă. Majoritatea abordărilor definesc conceptul de comunitate pe baza rezultatelor obținute în urma aplicării algoritmului de detecție. Astfel, nu putem ști neapărat dacă rezultatul obținut este unul corect, sau măcar relevant. În această lucrare, ne propunem să oferim o soluție care mai întâi definește în mod clar problema detectării comunităților și ce reprezintă o comunitate. Pe baza acestora, vom construi un model matematic pe care il vom testa mai apoi intr-o aplicatie si vom compara rezultatele obtinute cu alte solutii relevante din domeniu.

2 Incadrarea lucrării

ACM E.E1 Data - Data Structures - Graphs and Networks

AMS 05-XX 05-Cxx

3 Definirea conceptului de comunitate

Deoarece conceptul de comunitate este unul destul de greu de definit, putând lua multiple interpretări valide, vom încerca să plecăm de la un exemplu intuitiv și practic, urmând ca apoi să oferim și o definiție matematică riguroasă.

Din punct de vedere lingvistic, prin comunitate înțelegem un grup de oameni cu interese, aspirații, dorințe, trăsături comune, asemănătoare. Adică ceva sau cineva leagă, unește acești oameni. Evident, nu orice comunitate se bucură de aceleași legături. Unii membri sunt foarte implicați, poate își asumă chiar rolul de lideri, fiind foarte influenți, în schimb ce alții se mulțumesc cu statutul de simplu figurant. Putem privi problema atât din perspectiva membrului, cât și din cea a individului. Aceste perspective pot fi relativ identice, dar și destul de diferite, după cum putem deduce din exemplele următoare. Un membru se poate simți neintegrat și neîmplinit, în ciuda bunului mers al comunității(de exemplu, elevul retras la școală). Vice versa, un membru poate fi în relații bune cu comunitatea din care face parte,în ciuda disputelor interne(de exemplu, într-o familie numeroasă, în care pot apărea conflicte ușor, copiii sunt feriți, protejați și iubiți necondiționat de aproape toate rudele).

4 Pespectiva comunității

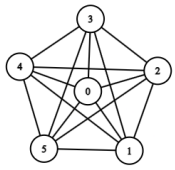
Principiul de funcționare al unei comunități este relația, legătura dintre membri. Această legătură poate fi de diferite tipuri: de prietenie, de colegialitate, de asociere, de colaborare, etnică, socială, religioasă, etc. Existența unei relații între indivizii unei comunități este modelată matematic de conceptul muchiei bidirecționale între două noduri dintr-un graf. Următoarea problemă care se pune este cum măsurăm, cum exprimăm din punct de vedere matematic cât de bine funcționează o comunitate. Din punct de vedere social, exprimat în termeni plastici, cazul ideal se obține atunci când "toată lumea se înțelege bine cu toată lumea", adică există relații de prietenie între toți membrii comunității. Privind înapoi la grafuri, acest lucru ar însemna că există o muchie bidirecțională între oricare două vârfuri. Adică avem de a face cu un graf complet.

DEFINIȚIE 1. O comunitate este un graf(sau subgraf) complet (fig1)

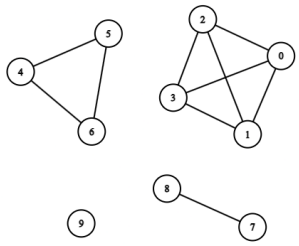
G = (V, E), $V = \bigcup_{i=1}^{n} i$, $E = \bigcup_{i=1}^{n} \bigcup_{j=1, i \neq j}^{n} (i, j)$

În majoritatea cazurilor însă, vor exista mai multe comunități. Putem extinde definiția anterioară pentru un caz mai general

DEFINIȚIE 2. Un graf poate fi partiționat în comunități dacă este o pădure de grafuri complete.(fig2)



Graf complet(comunitate)



Pădure de grafuri complete

Evident, definițiile prezentate anterior sunt foarte rigide și aproape imposibil de implementat în practică. Însă modelul matematic în caz ideal și analogia cu lumea reală constituie o bază solidă pentru atacarea problemei și identificarea unor formule de calcul cât mai relevante. La fel, vom face legături practice și intuitive care să poată conduce ulterior la un model matematic riguros. Având în vedere că situația ideală, de maximă legătură în cadrul unei comunități, este irelevantă, putem căuta un coeficient numeric care să ne indice cât de aproape, sau de departe este comunitatea noastră de acest obiectiv propus. Putem calcula ușor un procent de "umplere" a grafului, care se exprimă astfel: $\phi = \frac{|E|}{|V| \cdot (|V|+1)/2}$.

Evident, pentru un graf complet, $\phi = 1$, iar pentru un graf fără muchii, $\phi = 0$. Putem extinde problema și pentru cazul când vrem să măsurăm împărțirea unui graf în 2 sau mai multe comunități.

Considerăm următiarea partiție a grafului $G = (V, E)$:

$G = \bigcup_{i=1}^{k} G_i = \bigcup_{i=1}^{k} (V_i, E_i)$

Să explicăm aceste notații:

G_i = graful asociat comunității

V_i = vârfurile din graful G_i, adică membrii comunității i

E_i = muchiile între două vârfuri din G_i Atenție - În E_i nu sunt incluse muchiile între noduri din comunitatea i și noduri din alte comunități. Acest aspect este foarte important și va fi tratat ulterior.

Observații:

$\bigcup_{i=1}^{k} V_i = V$, adică fiecare nod trebuie să aparțină unei comunități.

$\cap_{i=1}^{k} V_i = \emptyset$, adică un nod nu poate aparține de mai multe comunități.

Combinând cele două observații de mai sus, ajungem la concluzia că orice nod aparține exact unei comunități.

$\cap_{i=1}^{k} E_i = \emptyset$

$\bigcup_{i=1}^{k} E_i \subseteq E$

$E - \bigcup_{i=1}^{k} E_i = \{(x, y) \mid (x, y) \in E, x \in V_{i1}, y \in V_{i2}, i_1 \neq i_2\}$

O idee trivială pentru a calcula coeficientul ϕ întregului graf este de a calcula valoarea lui ϕ pentru fiecare comunitate în parte, și a face media aritmetică a tuturor acestor valori. Această abordare are însă un mare dezavantaj, și anume că nu ia în niciun fel în calcul muchiile extra-comunitare. Astfel, putem obține multe comunități micuțe, cu o modularitate internă ridicată, dar a căror membri sunt strâns legați și de alte comunități. Practic, algoritmul ar descompune o comunitate mai mare, cu un coeficient ϕ scăzut, în multe sub-comunități de dimensiuni reduse, dar cu un coeficient ϕ mai bun. Cazul cel mai rău ar fi cel în care fiecare nod alcătuiește de unul singur o comunitate, care evident are $\phi = 1$. O posibilă soluție este să includem în calcul numărul muchiilor extra-comunitare, care cu cât este mai mic, cu atât rezultatele sunt mai relevante și oferă o imagine corectă a realității. Aceasta ar putea fi exprimată procentual, privind raportul dintre numărul muchiilor intra-comunitare și numărul total al muchiilor din graf. Putem exprima coeficientul ϕ pentru întregul graf astfel:

$\phi_G = \frac{\sum_{i=1}^{k} \phi_i}{k} \cdot \frac{|\bigcup_{i=1}^{k} E_i|}{|E|}$

5 Perspectiva membrului

Până acum am încercat să exprimăm și să măsurăm partiția unui graf în comunități privind exclusiv la optimalitatea fiecărei comunități în parte. În cele ce urmează, vom acorda o atenție sporită indivizilor, fără a ne interesa neapărat comunitatea din care fac parte ca ansamblu. Privind la ceea ce se întâmplă în realitate, putem observa că se ridică următoarea problemă. În viața de zi cu zi, fiecare face parte din mai multe anturaje, grupuri, comunități - în familia restrânsă, în familia lărgită, la școală sau la locul de muncă, cu prietenii, la practicarea unui hobby comun etc.Cu toate acestea, nu ne simțim la fel de atașați de toate comunitățile din care facem parte, și nu toate comunitățile au aceeași relevanță, influență. Dacă ar fi să ne gandim cu care comunitate ne identificăm cel mai mult, probabil ar fi cea cu care avem cea mai strânsă legătură. putem descrie cât de bine este integrat în comunitatea respectivă judecând după numărul de asocieri cu ceilalți membri. De exemplu, intr-o clasă, putem observa cu câți dintre colegisăi este prieten un elev. Problema pe care trebuie să o prevenim la fel ca în cazul anterior, este de a nu izola nodul de restul comuniăților. Trebuiesă privim și la cât de relevante sunt aceste relații din comunitatea proprie. De exemplu, într-un anturaj restrâns, o persoană poate fi prieten cu toți membrii acelui subgrup, dar să aibă mult mai multe relații de prietenie cu persoane din alte grupuri, acest lucru indicând că deși estefoarte bine integrat în comunitatea inițială, nu este neapărat cea mai relevantă pentru el. Matematic, putem exprima ideea printr-un număr cât mai ridicat de muchii ale unui nod cu alte noduri din comunitatea de care aparține, și un număr cât mai scăzut de muchii cu noduri din alte comunități. Cazul ideal este când individul nu are legături ex-comunitare, și este strâns legat de comunitatea lui.

$\lambda_i = \frac{card\{(x,y)|i\in V_p,(x,y)\in E_p\}}{|V_p|} - \frac{card\{(x,y)|i\notin V_p,(x,y)\in E_p\}}{card\{(i,j)|(i,j)\in E)\}}$ Să explicăm această formulă:

Fracția din stânga reprezintă raportul dintre numărul de muchii intra-comunitare ale nodului i, și dimensiunea comunității i, adică raportul dintre câți prieteni are individul în comunitatea respectivă și câți poate avea în total, sau, mai simplu, un număr subunitar care arată cât de bine este integrată persoana.

Fracția din dreapta reprezintă raportul dintre numărul de muchii extra-comunitare ale nodului i și numărul total de muchii, adică raportul dintre câți prieteni are individul în alte comunități și câți are în total în orice comunitate, sau, mai simplu, un număr subunitar care arată gradul de nefidelitate a persoanei față de propria comunitate.

Cel de-al doilea număr subunitar rezolvă astfel problema descrisă anterior, un scor ridicat la nivelul propriei comunități pentru nodul respectiv fiind anulat de operația de substracție.

Pentru întregul graf, vom măsura coefientul λ astfel $\lambda_G = \frac{\sum_{i=1}^{|V|}\lambda_i}{|V|}$

6 Implementare

Unul dintre cele mai populari algoritmi de detectare a comunităților într-un graf este algoritmul Girvan-Newman[4], bazat pe conceptul de modularitate. Algoritmul este unul de tip greedy, căutând la fiecare pas să maximizeze creșterea modularității, având drept condiție de oprire atingerea unui parametru pentru numărul de comunități dorit, sau că nu se mai pot face îmbunătățiri la valoarea modularității. La început, fiecare nod este o comunitate de sine stătătoare, urmând să unim comunitățile care vor genera cea mai mare creștere a modularității. Diferența în implementarea noastră ar fi că în loc de modularitate, vom căuta să maximizăm coeficienții ϕ_G, respectiv λ_G. Acest algoritm are destule limitări, rezultatele obținute fiind neconcludente(anumite join-uri de comunități pot fi foarte bune pe moment, dar ar putea conduce la niște valori de maxim local, algoritmul urmând să se plafoneze rapid). Această problemă ar putea fi tratată prin aplicarea unor operații de undo asupra join-urilor anterioare, pentru a scăpa din capcana maximului local). Un algoritm de backtracking ar rezolva această problemă, însă este foarte limitat pe date reale de dimensiuni mari. Astfel, soluția agreată pentru a trata atât problema eficienței ca timp de execuție, cât și a maximului local, este implementarea unui algoritm evolutiv de tip nedeterminist.[16]. Ca funcție de fitness, vom folosi parametrii ϕ, respectiv λ.Deoarece parametrii obținuți nu ne spun multe, vom măsura modularitatea soluției obținute, și o vom compara cu soluția care folosește ca funcție de fitness modularitatea. Se vor efectua următoarele experimente:

1. Pornind de la un graf oarecare, să se determine o partiție a grafului în comunități, care maximizează modularitatea/ϕ/λ.(câte o execuție separată pentru fiecare parametru)

2. Să se calculeze modularitatea pentru partiția obținută la punctul 1, pe baza lui /ϕ/λ. Această valoare se compară cu valoarea obținută la pasul 1 pentru modularitate.

7 Descrierea algoritmului evolutiv

Algoritmii evolutivi se diferențiază de algoritmii clasici prin faptul că sunt ne-determiniști, adică la multiple execuții ale programului, se vor obține soluții diferite. Am explicat deja în secțiunea anterioară avantajele acestei abordări. În cele ce urmează, vom prezenta particularitățile acestui algoritm evolutiv.

I. Genotipul

Articolul citat propune mai multe variante pentru codificarea genetică a unei soluții. Metoda aleasă este locus-based.

$neighbors = \bigcup_{i=1}^{|V|} j, (i,j) \in E$

$neighbors_i$ reprezintă un vecin al lui i care se află în aceeași comunitate cu i

Am ales această reprezentare, luând în calcul următoarele aspecte:

1. Spațiul de căutare al soluției se reduce de la $|V|^{|V|}$ la $\prod_{i=1}^{|E|} deg_i$ (deg_i=gradul nodului i), astfel algoritmul evolutiv converge spre punctul de optim global mai rapid.

2. Această abordare facilitează detectarea comunităților. Pe baza șirului $neighbors$, putem construi un subgraf al grafului inițial(citit din fișier)

$G' = (V, E'), E' = \{(i,j) \mid neighbors_i = j \vee neighbors_j = i\}$

Problema detectării comunităților se transformă într-o simplă problemă de detectare a componentelor conexe ale grafului G'.

II. Prima generație

Pentru prima generație, vom alege în mod aleator, o valoare pentru $neighbors_i$, dintre nodurile cu care are muchie.

III. Mutația

Mutația are rolul de a conferi diversitate în rândul indivizilor. În problema noastră, se alege aleator un nod i al unei soluții, pentru care, tot aleator, alegem un alt vecin j pentru $neighbors_i$. Deși soluția pare banală, se dovedește a fi foarte eficientă. Comunitatea lui i fie va scinda în două comunități mai mici sau se va uni cu alta.

IV. Încrucișarea

Se creează aleator o mască binară de lungime $|V|$, pentru a indică părintele de la care se alege alela j.

	1	2	3	4	5	6	7	8	9	10	11	12
Parent 1	3	1	4	7	4	12	4	7	8	9	12	6
Parent 2	2	9	2	5	6	7	8	10	2	11	10	8
Mask	1	1	0	0	1	1	0	0	0	1	1	1
Child	2	9	4	7	6	7	4	7	8	11	10	8

Exemplu de încrucișare

8 Testare

Pentru partea de testare, vom folosi în primă fază niște date de dimensiuni reduse, care pot fi ușor vizualizate în mediul de dezvoltare, Apoi, vom testa și pe date de dimensiuni ridicate, folosind input-uri oferite de Newman (krebs, football, dolhpins, karate, lesmiserables).

	ϕ	λ	**mod**	ϕ **+ mod**	λ **+ mod**
krebs	0.11	0.25	0.52	0.35	0.47
football	0.05	0.40	0.58	0.33	0.56
dolphins	0.14	0.33	0.53	0.28	0.45
karate	0.13	0.34	0.42	0.18	0.35
lesmiserables	0.12	0.50	0.56	0.25	0.50

Rezultate

ϕ, λ, mod - valorile obținute în urma efectuării primului experiment

ϕ + mod, λ + mod - valorile obținute în urma efectuării celui de-al doilea experiment

Ultimele 3 coloane sunt cele relevante și pe care le folosim pentru a compara parametrii și a trage concluzii.

9 Bibliografie

1. Goldenberg, D. (2021). Social network analysis: From graph theory to applications with python.arXiv preprint arXiv:2102.10014.

CAMPBELL, William M.; DAGLI, Charlie K.; WEINSTEIN, Clifford J. Social network analysis with content and graphs.Lincoln Laboratory Journal, 2013, 20.1: 61-81.

2. Leung, I. X., Hui, P., Lio, P., Crowcroft, J. (2009). Towards real-time community detection in large networks. Physical Review E, 79(6), 066107.

3. Gulbahce, N., Lehmann, S. (2008). The art of community detection. BioEssays, 30(10), 934-938.

4. NEWMAN, Mark EJ. Fast algorithm for detecting community structure in networks. Physical review E, 2004, 69.6: 066133.

5. Santo Fortunato, Darko Hric,Community detection in networks: A user guide,Physics Reports, Volume 659,2016, Pages 1-44,ISSN 0370-1573

6. J. Yang, J. McAuley and J. Leskovec, "Community Detection in Networks with Node Attributes," 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, 2013, pp. 1151-1156, doi: 10.1109/ICDM.2013.167.

7. XIE, Jierui; KELLEY, Stephen; SZYMANSKI, Boleslaw K. Overlapping community detection in networks: The state-of-the-art and comparative study.Acm computing surveys (csur), 2013, 45.4: 1-35.

8. Lancichinetti, A., Fortunato, S., Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. Physical review E,78(4), 046110.

9. JIN, Di, et al. A survey of community detection approaches: From statistical modeling to deep learning. IEEE Transactions on Knowledge and Data Engineering, 2021, 35.2: 1149-1170.

10. RUAN, Yiye; FUHRY, David; PARTHASARATHY, Srinivasan. Efficient community detection in large networks using content and links. In:Proceedings of the 22nd international conference on World Wide Web. 2013. p. 1089-1098.

11. Traag, V. A., Bruggeman, J. (2009). Community detection in networks with positive and negative links. Physical Review E,80(3), 036115.

12. Orman, G. K., Labatut, V., Cherifi, H. (2012). Comparative evaluation of community detection algorithms: a topological approach.Journal of Statistical Mechanics: Theory and Experiment, 2012(08), P08001.

13. Hastings, M. B. (2006). Community detection as an inference problem. Physical Review E, 74(3), 035102.

14. Peel, L., Larremore, D. B., Clauset, A. (2017). The ground truth about metadata and community detection in networks.Science advances,3(5), e1602548.

15. PIZZUTI, Clara. Evolutionary computation for community detection in networks: A review IEEE Transactions on Evolutionary Computation, 2017, 22.3: 464-483.