Xarxes Socials i Econòmiques
2a part Examen Final

Maria del Mar Bibiloni

MADM, UIB. Curs 2017-18

Read the paper "Social triangles and generalized clustering coefficient for weighted networks" by R. Cerqueti, G. Ferraro and A. Iovanella (arXiv preprint arXiv:1712.01561 (5 Dec. 2017), https://arxiv.org/abs/1712.01561), and then answer the following questions.

1) *In a single sentence, what is the main question addressed in this paper? Why do the authors consider this question relevant and worth answering? After reading the paper, do you consider their answer relevant? Why?*

La qüestió que tracta l'article és com analitzar les arestes *indirectes* d'un graf amb pesos.

Aquestes arestes *indirectes* sorgeixen quan tenim tres nodes, v_1, v_2, v_3, dues arestes v_1v_2, v_2v_3, i pesos w_{12}, w_{23} molt grans. La idea que sorgeix és que els nodes v_1 i v_3 estan tan fortament connectats per v_2 que no es fa necessària una aresta directe entre v_1v_3. El fet que no existeixi aquesta aresta, fa que no es consideri cap triangle amb els tres nodes per vèrtexs, però si pensam amb lo *fort connectats* que estan els tres nodes, sí s'hauria de considerar.

Concretem més amb l'exemple que es deriva dels comentaris a la pàgina 3 de l'article. Suposem que tenim un graf estrella, i el node central està connectat amb tots els altres amb un pes molt elevat. Donats dos nodes v_i, v_j qualsevols que no són el central, tenim que no hi ha aresta entre ells, per tant, no hi ha assortativitat, és a dir, els nodes de mateix grau ($d_i = 1$) no tendeixen a connectar-se. Aquest exemple però, admet una altra interpretació: tots aquests nodes estan altament connectats pel node central. En altres paraules, existeix una aresta *indirecta* entre cada parell de nodes no centrals. Així, sí podríem considerar la assortativitat de la xarxa.

A més, hi pot haver triangles on una de les arestes tingui un pes molt petit i, per tant, aquesta aresta sigui irrellevant. Aquest fet fa que, si aquest triangle es compta com a tal, també s'hauria de comptar un triangle sense dita aresta.

Tot això fa que els autors considerin necessari un nou concepte de triangles per aquestes xarxes i, com a conseqüència, una nova definició de coeficient de clustering.

La pregunta dels autors és rellevant perquè pot ajudar a explicar millor cert tipus de xarxes. Així i tot, és un coeficient molt específic que no sempre funciona. Per exemple, considerem un graf d'amistats on els pesos indiquen el nombre de vegades que dos nodes han mantingut una conversació. En aquest cas, és cert que si dos nodes han parlat només un pic entre ells i molt amb un v_i, tenen gairebé el mateix comportament que si mai haguessin parlat. Així i tot, no seria convenient considerar-ho com un triangle.

2) *How does this paper relate to previous work about the question addressed in it?*

Per un banda, l'article presenta altres propostes de coeficients de clusterig per a xarxes amb pesos a les arestes. Tots aquests coeficients tenen una cosa en comú, només es calculen a partir dels pesos de triangles. D'aquesta manera, ningun permet arestes *indirectes* en el graf. Així, es recolza en aquest fet per introduir la seva definició.

D'altra banda, revisa els estudis fets anteriorment per reforçar la idea que el coeficient de clustering té un paper rellevant a l'hora d'analitzar l'estructura d'una xarxa. Per exemple, ens parla de que es pot emprar per detectar *small-worlds* i mesurar la connectivitat entre grups de veïns. D'aquesta manera, pretén fer veure que aquests conceptes podrien ser flexibles a arestes *indirectes* segons la xarxa i, per tant, també ho ha de ser la definició.

3) *What are the main definitions in this paper?*

Les definicions principals de l'article són la definició de pseudo-triangles i la nova definició de clustering que empra aquests triangles.

Dels tres tipus de triangles que defineix l'article, cal destacar que es determina el valor dels pesos pels quals cal considerar una aresta *indirecta* segons una funció F. Encara que es donin exemples de funcions F que ells empren, el fet de d'imposar poques condicions sobre F fa que el seu model sigui flexible a molts tipus de xarxes, ja que es podrien arribar a definir tot tipus d'arestes indirectes.

Així, aquesta nova definició de triangle esdevé la principal, més que la de coeficient de clustering, que tot i ser nova no s'allunya tant de la definició clàssica.

4) *On your behalf, what are the strongest contribution and the greatest drawback of this paper (if any)?*

Seguint amb la resposta a la pregunta **3)**, la major contribució dels autors de l'article és pensar en que hi ha xarxes on s'ha de canviar de noció de triangle per entendre la connectivitat entre els nodes. Així, pot haver-hi aquestes connexions fortes existents que no es tradueixin en aresta, però que fan que una tripla s'hagi de considerar un triangle per explicar millor la estructura de la xarxa. D'aquesta manera, malgrat pugui haver millors maneres de definir el coeficient de clustering, la clau està en considerar un nou ventall de concepte de pseudo-triangles.

D'altra banda, sorgeixen problemes amb els triangles de T_3, on el node i és l'extrem d'un camí de distància dos. Així, segons la seva definició i els exemples, sembla que si es restringeix el nombre de triangles T_3, el valor de $C_i^{(g)}$ esdevé molt petit i pareix que la xarxa està menys connectada a través de triangles i pseudo-triangles. L'objectiu però, entenc que és considerar els pseudo-triangles per veure que el coeficient de clustering generalitzat és major que l'usual, és a dir, el graf està més connectat del que es detectava.

A més, malgrat els exemples, no es fa notar que prendre $\alpha = 0$ i $\beta = 0$ alhora no es útil quan s'ha de calcular un coeficient de clustering, segons el nombre de triangles i pseudo-triangles, ja que com tota tripla conta com a triangle el coeficient és 1, però tota tripla no s'hauria de considerar com un triangle *indirecte*. Per exemple, un graf anell no té sentit que tengui coeficient de clustering 1.

5) *Do you consider the numerical experiments provided in Section 5 suitable? (I'm not referring to the networks analyzed, but to the computations performed on them.) What conclusions do you extract from these experiments? Do they agree with the authors' conclusions? What other experiments (on the same networks) would you perform?*

Els experiments realitzats a l'article, recolzen la idea de considerar pseudo-triangles per donar importància a arestes *indirectes*, ja que pot canviar la manera d'entendre l'estructura de la xarxa. Així, per exemple, el coeficient de clustering mitjà usual de la xarxa C.elegans és $\bar{C} = 0.228$, mentre que pel valor de $\alpha = 10$ i $\beta = 4$ tenim $C^{(g)} > 0.3$ per F_1 i F_2. Per F_1, de fet, $C^{(g)} > 0.5$. Tinguent en compte que la mitjana de pesos és 4.198, aquests valors de α i β són acceptables. Aquestes dades fan referència a la Figura 11.

El segon exemple(US airport), encanvi, no ens aporta gaire informació sobre la importància d'emprar pseudo-triangles. Així, $\bar{C} = 0.617$, mentre que a la Figura 12 tenim que pel valor de $\alpha = 10$ i $\beta = 0$, $C^{(g)} > 0.8$. Pot semblar un coeficient bo, però fixem-nos amb els valors de α i β. Per un costat, considerar només fins α, $\beta = 10$ és molt poc quan la mitjana de pesos és 152320.19. Així, en aquest cas, s'hauria d'haver estes l'experiment a molts més valors, sobretot de α, que sembla mantenir-se més estable el coeficient de clustering $C^{(g)}$. Almenys, hagués estat interessant considerar fins $\alpha = 152320.19$.

Una de les conclusions que treuen dels experiments, és que els triangles de tipus T_2 tenen una tendència a tenir un pes bastant alt. Aquesta conclusió es recolza en el fet que augmentant el valor de α el coeficient de clustering no varia gaire. Per tant, hi ha xarxes on pot ser rellevant considerar els triangles de tipus T_2. Com s'ha comentat abans, sembla haver un problema a l'hora de considerar els triangles de tipus T_3 de manera restringida (β gran).

6) *In their experiments, the authors mostly consider cases when $\beta = 0$ or $\alpha = 0$. How would you choose α and β so that only really meaningful pseudo-triangles were considered?*

Per tal de considerar pseu-triangles que realment es puguin interpretar com "la tercera aresta és irrellevant perquè els pesos de les altres són suficientment grans" i, per tant, tenim una aresta *indirecta*, hem de prendre α i β de manera que es satisfaci la condició. Així, si la mitjana de pesos és $\bar{w}$ jo empraria $F_2(w_{ij}, w_{ik}) = \frac{w_{ij}+w_{ik}}{2}$ i consideraria $\alpha = \beta = \bar{w}$. La interpretació seria, si la mitjana de pesos del candidat a pseudo-triangle supera la mitjana de pesos del graf, aleshores l'aresta que falta és innecessària.

La funció a emprar i el valor de α i β poden variar en funció de quina xarxa tractam, com estiguin distribuïts els pesos, etc. Però en qualsevol cas no consideraria $\alpha = 0$ o $\beta = 0$, ja que estaríem dient que totes les triples o tots els camis de distància 2 (partint d'un node) són triangles.

7) *The authors define their generalized clustering coefficient only for undirected networks. How would you generalize it in a sound way to directed networks? Provide a definition that makes sense on some class of weighted directed networks of your choice with some specific meaning of its links (for instance, non-symmetric friendship networks, trade networks, citation networks with weights the number of times an author has cited another author, . . .)*

Per crear pseudo-triangles que tinguin sentit per grafs dirigits, recordem quina és la idea de pseudo-triangle: si existeix ij i ik amb pesos suficientment grans, jk és innecessària. Així, anem a interpretar com a aresta innecessària una aresta tal que no faci falta perquè arribi la informació de j a k ràpidament i anem a definit els nostres triangles.

Definició 1. *Considerem un graf dirigit amb pesos $G = (V, E)$ amb nodes $V = \{1, .., N\}$, matriu d'adjacència $A = (a_{ij})i, j = 1, ..., N$ i matriu de pesos $W = (w_{ij})_{i,j=1,...,N}$, amb pesos no negatius. A més, prenem $\alpha, \beta \in [0, \infty)$ i una funció $F : [0, +\infty)^2 \to [0, +\infty)$ no decreixent.*

Per a cada $(i, j, k) \in V^3$, un subgraf $t = (i, j, k, E_T)$ és un triangle que de i si es satisfan alguna de les següents condicions.

$\widetilde{T}_1$ $a_{ij} = 1$, $a_{ik} = 1$, *i* $a_{jk} = 1$ *o* $a_{kj} = 1$;

$\widetilde{T}_2$ $a_{ij} = 1$, $a_{ik} = 1$, $a_{jk} = a_{kj} = 0$ *i* $F(w_{ij}, w_{ik}) \geq \alpha$;

$\widetilde{T}_3$ $a_{ij} = 1$, $a_{ki} = 1$, $a_{jk} = a_{kj} = 0$ *i* $F(w_{ij}, w_{ik}) \geq \beta$.

D'aquesta manera, tenim que els triangles de tipus T_1 han de tenir la forma de les triples de la Figura 1 marcades com **Sí** (que són anàlogues) i la tercera aresta pot prendre qualsevol direcció. D'altra banda, els triangles de Tipus T_2 són aquells de la forma **Sí**, sense l'aresta faltant. La idea és que almenys un dels nodes adjacents pot passar la informació ràpidament a l'altre, per tant, no necessita una aresta més. Finalment, els triangles de tipus tres serien camins de longitud 2, $i \longrightarrow j \longrightarrow k$, de manera que si i pot passar informació ràpidament a k no necessita un camí directe.

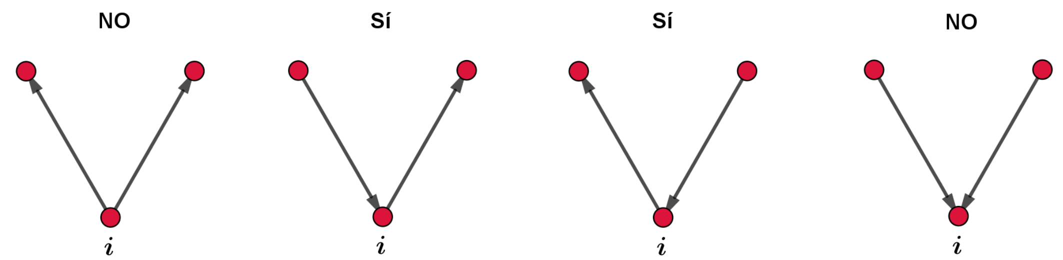

Figura 1: Triangles de tipus T_2, si $F(w_{ij}, w_{ik}) \geq \alpha$.

Ara, anem a definir el coeficient de clustering de la manera usual: (nombre triangles existents)/(nombre possibles triangles) + $C_i^{(g)}$.

Definició 2. *Donat un graf $G = (V, E)$ dirigit amb pesos i un node $i \in V$, el coeficient de clustering de i es*

$$\widetilde{C}_i^{(g)} = \frac{|\ \widetilde{T}_1 \cup \widetilde{T}_2 \cup \widetilde{T}_3\ |}{\widetilde{D}_i} + D_i^{(g)},$$

on $\widetilde{D}_i = degree_{in}(i) degree_{out}(i) + |j \in V : \exists\ i \to k \to j, \text{per algún } k \text{ i } \nexists (i,k)|.$

9) *How is the generalized clustering coefficient related to the Strong Triadic Closure Property? Can you find a way to test whether a network satisfies the STCP using this coefficient (together with other indices, of course)?*

La propietat STCP ens diu que si existeixen arestes ij i jk fortes, aleshores existeix una aresta entre i i k. Aquí, ens parla d'arestes existents en el graf, no d'arestes *indirectes*. Per tant, no és adient emprar el coeficient de clustering generalitzat a un graf que satisfà STCP; les arestes *indirectes* existeixen en el graf i, sinó, és perquè les altres arestes no eren suficientment fortes.

De fet, si un graf satisfà STCP, vol dir que no tindrà quasi triangles de tipus T_2 ni T_3, ja que donats dos nodes a distància dos, hi ha una forta tendència a que existeixi una aresta que redueix la distància a 1 i la majoria de possibles T_2 seran en realitat tancats. En aquest cas, tindrem que el coeficient de clustering generalitzat i el coeficient usual seran quasi iguals, prenent un valor pròxim a 1.

En resum, calcular els dos coeficients i veure si es semblen podria ser una manera per comprovar si una xarxa satisfà STCP.

10) *In page 4 the authors say "Our generalized clustering coefficient has a further very relevant property: it assumes unitary value in several situations and not only when the graph is a clique." Is it true that if an unweighted graph has clustering coefficient 1, then it is a clique? And if we assume moreover that the graph is connected? Justify your answer.*

- *Si un graf sense pesos té coeficient 1, aleshores és un clique.*
  FALS. Basta considerar el graf de la Figura 2, composat per un triangle i un node de grau zero.

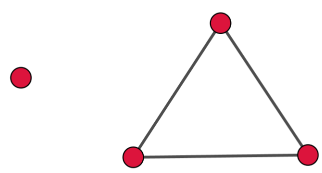

Figura 2: Graf sense pesos amb coeficient de clustering 1 i no complet.

Tots els nodes del triangle tenen coeficient de clustering 1, mentre que el node aïllat té coeficient 0. Quan es fa la mitjana dels coeficients, només es tenen en compte aquells de grau major que 1, per tant, el node aïllat no compta i el coeficient de clustering mitjà resulta 1.

- *Si un graf connex sense pesos té coeficient 1, aleshores és un clique.* CERT.

  Demostració. Considerem $G(V, E)$ un graf connex sense pesos i amb coeficient de clustering 1.

  Si el coeficient de clustering és 1, almenys ha d'existir un triangle (altament seria zero). Aleshores, $|E| \geq 3$.

  Tenint en compte aquest fet, anem a demostrar el resultat per inducció en el nombre de vèrtexs $|V|$.

  - Cas inicial. Considerem un graf de 3 nodes i 3 arestes. Donat que té coefcient de clusterig 1 ha d'existir un triangle, per tant, ha de ser un triangle. Així, tenim que és un clique.

  - Pas d'inducció. Suposem que si $|V| = n$, el coeficient de clustering $c = 1$ i el graf és connex, aleshores és un clique.

    Ara, afegim un nou node v_* al clique i imposam que $G_* = G(V \cup \{v_*\})$ és connex i $c = 1$.

    Dotat que G_* es connex existeix l'aresta v_*v_i per algun v_i de G. A més, existeix v_iv_j per a tot v_j de G, ja que és un clique. En altres paraules, v_i està connectat amb tothom de G.

    Ara, com existeix v_*v_i i v_iv_j, ha d'existir v_*v_j per a tot v_j; sinó hi hauria almenys una tripla més que triangle i $c \neq 1$. En resum, v_* està connectat amb tots els nodes de G i tots els nodes de G estan connectats entre ells, aleshores, G_* és un clique.

  □

11) *In equation (3), page 7, the authors recall Barrat et al's clustering coefficient for weighted networks. I also introduced this index in my lectures, but equation (3) and my definition do not define the same number. What is the difference? Which one corresponds to Barrat et al's original definition? How did you check it? If the correct version is the one given in this paper, I probably had a good reason to change it. What was this reason? And, if the correct version is the one I gave in the lectures, did the authors of this paper have a good reason to change it (for instance, to compare it with their proposed coefficient)?*

Segons l'article, el coeficient de clustering per grafs amb pesos introduït per Barrat el al. és el següent.

$$\widetilde{C}_{i,\mathrm{B}} = \frac{1}{s_i(d_i - 1)} \sum_{j,k \in V} \frac{w_{ij} + w_{ik}}{2} a_{ij} a_{jk} a_{ik}, \tag{1}$$

on w_{ij} és el pes de l'aresta ij, $s_i = \sum_j w_{ij}$, $d_i = \text{degree}(i)$ i a_{ij} és l'element (i,j) de la matriu d'adjacència del graf, és a dir, val 1 si existeix l'aresta ij, 0 altrament.

D'altra banda, segons la definició donada a classe tenim

$$\widetilde{C}_{i,\mathrm{B}} = \frac{1}{s_i(d_i-1)} \sum_{i,j,k \text{ triangle}} (w_{ij} + w_{ik}). \tag{2}$$

Si cercam l'article de Barrat *et al.*, http://www.pnas.org/content/pnas/101/11/3747.full.pdf, trobam la definició (1). Ambdues però són equivalents.

Per començar, la primera diferència que s'observa és que a la suma de (1) apareixen els productes $a_{ij}a_{jk}a_{ik}$. Així, l'únic que es fa és sumar només els pesos d'arestes que pertanyen a triangles. Per tant, sembla que la diferència està en què a (1) tenim $\frac{w_{ij}+w_{ik}}{2}$ i a (2) $w_{ij}+w_{ik}$. Encara que ho pugui semblar, en el cas (1) no es vol calcular la mitjana dels pesos, sinó que es divideix entre dos perquè es fa la mateixa suma dos cops. És a dir, es suma sobre $j, k \in V$ i, per tant, sumam $w_{ij}+w_{ik}$ i $w_{ik}+w_{ij}$; però ikj i ijk són el mateix triangle.

D'aquesta manera, als apunts de classe s'ha hagut de modificar la definició perquè es suma directament sobre els triangles. Així, només es realitza la suma $w_{ij}+w_{ik}$ una única vegada.