Centrality Algorithms (Algoritmi de centralitate)

- Algoritmii de centralitate sunt utilizați pentru a înțelege rolurile anumitor noduri într-un graf și impactul lor asupra rețelei respective.
- Sunt utili deoarece identifică cele mai importante noduri și ne ajută să înțelegem dinamica grupului, cum ar fi credibilitatea, accesibilitatea, viteza cu care lucrurile se răspândesc și legăturile dintre grupuri.
- Deși multi algoritmi dintre aceștia au fost inventați pentru analiza rețelelor sociale, de atunci s-au găsit utilizări în diverse industrii si domenii.

Se vor studia următorii algoritmi:

1. **Degree Centrality,** Gradul de centralitate ca masură de bază a conexiunii
2. **Closeness Centrality**, apropierea de Centralitate măsoară cât de central este un nod în grup, inclusiv două variante pentru grupurile neconectate.
3. **Betweenness Centrality**, pentru gasirea punctelor de control, include și o alternativă la aproximare.
4. **PageRank**, pentru înțelegerea influenței globale; include opțiuni populare de personalizare.

Centrality Algorithms (Algoritmi de centralitate)

- Majoritatea algoritmilor de centralitate calculează cele mai scurte căi între perechi de noduri. Pentru grafuri de dimensiuni mici și mijlocii funcționează fluent, dar pentru grafuri mari poate fi computațional dificil. Pentru a se evita durate lungi de rulare pe grafuri mari, unii algoritmi (de exemplu, Betweenness Centrality) au versiuni aproximative.
- Fiecare algoritm este inclus în Tabel, cu o scurtă descriere a algoritmului și informații despre funcționare. Versiuni ale algoritmilor deja tratați vor include mai puține detalii.

Tip algoritm	Descriere	Exemple
1) Degree Centrality	Măsoară numărul de relații ale unui nod	Estimarea popularității unei persoane pe baza gradului in-degree, de intrare in nod și out-degree pentru estimarea gregarității (grad de apartenență la un grup) (gregariousness)
2) Closeness Centrality Variante: **Wasserman și Faust, Harmonic Centrality**	Calculează care noduri au cel mai scurt drum (shortest paths) la toate celelate noduri	Găsirea locației optime pentru un nou serviciu public cu maximă accesibiltate
3) Betweenness Centrality Variante: Randomized Approximate Brandes	Măsoară numărul celor mai scurte drumuri (shortest paths) ce traversează un nod	Imbunătățirea țintelor pentru medicamente prin găsirea genelor ce controlează o anumită afecțiune medicală.
4) PageRank Varianta: **Personalized PageRank**	Estimează importanța nodului curent din legătura sa cu vecinii săi și a vecinilor lor (propus de Google)	Găsirea celor mai influente caracteristici ce pot fi extrase in Machine Learning și clasificare text pe bază de relevanță in Procesarea Naturală a Limbajului (NLP)

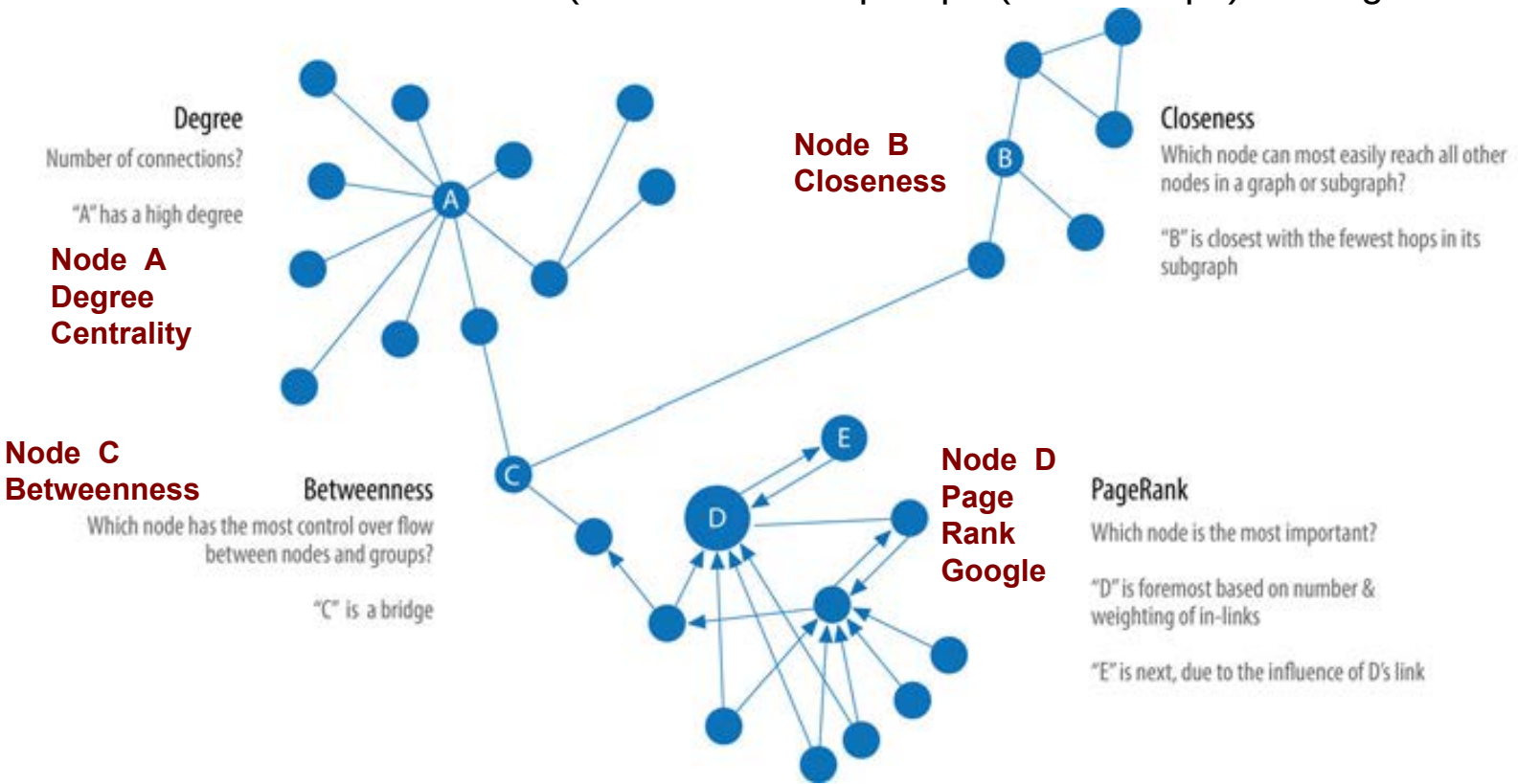
Exemplu de date Graful social

Algoritmii de centralitate sunt relevanți pentru toate grafurile, dar rețelele sociale oferă modalitatea potrivita de a gândi influența dinamica și fluxul de informații.
Exemplele sunt executate pe un graf similar Twitter. Ce tipuri de întrebări apar?

1) **Degree Centrality,** Gradul de centralitate ca măsură de bază a conexiunii
Ce număr de conexiuni există? (ex. nod A are gradul cel mai mare, nr. Max. conexiuni)

---

2) **Closeness Centrality**, Apropierea de Centralitate-măsoara cât de central este un nod în grup
De la ce nod se poate ajunge mai rapid la **toate** nodurile din graf sau subgraf?
(nod B cel mai aproape (fewest hops) in subgraful său)



Exemplu de date Graful social

Algoritmii de centralitate

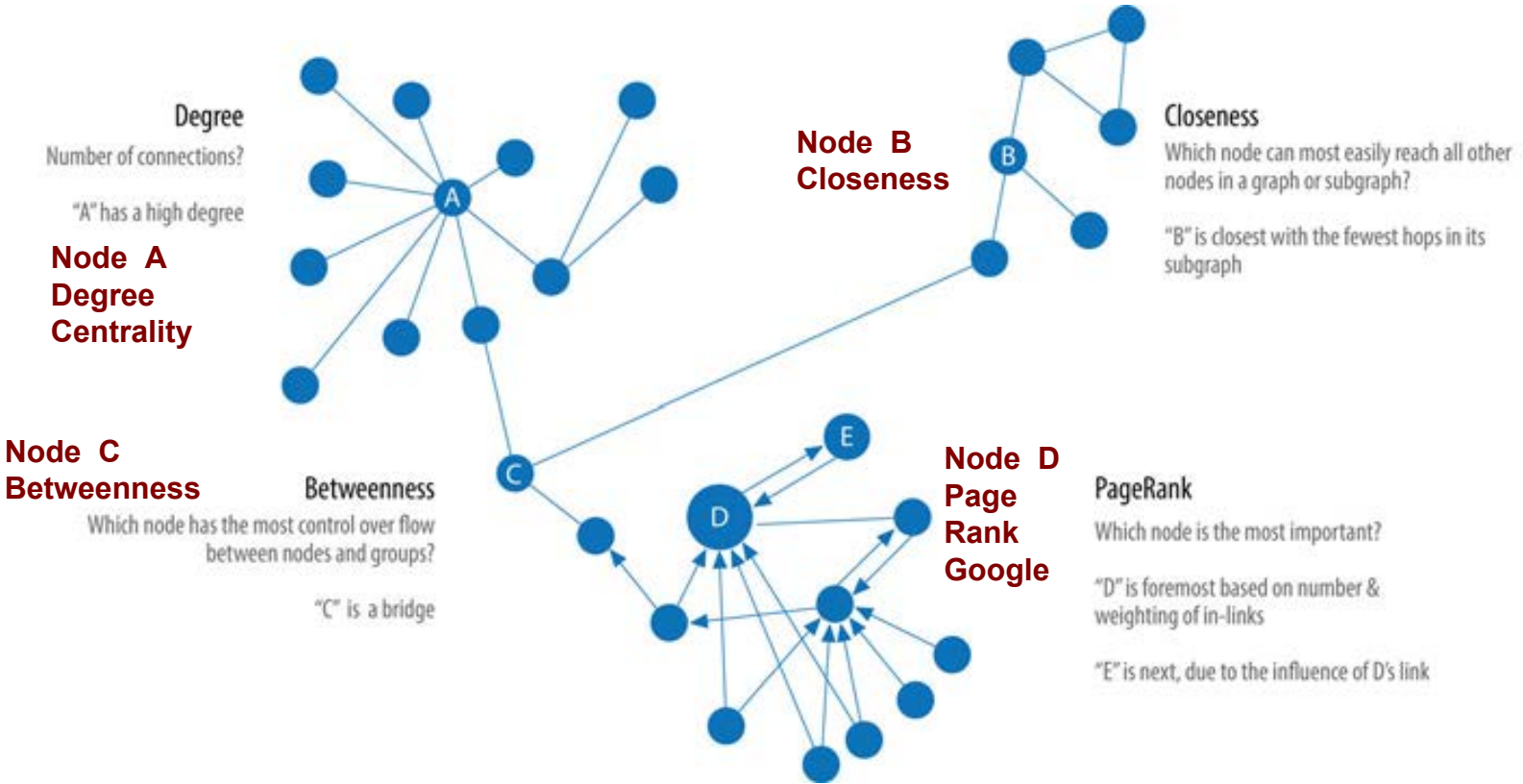
3) **Betweenness Centrality -** găsirea punctelor de control

Ce nod detine controlul major asupra relatiilor (flow) dintre noduri si grupuri?

(ex. nod C “bridge”)

4) **PageRank -** înțelegerea influenței globale cu opțiuni populare de personalizare.

Ce nod este cel mai important ? (nod D (nr. si cost relații cu celelate noduri), urmat de E



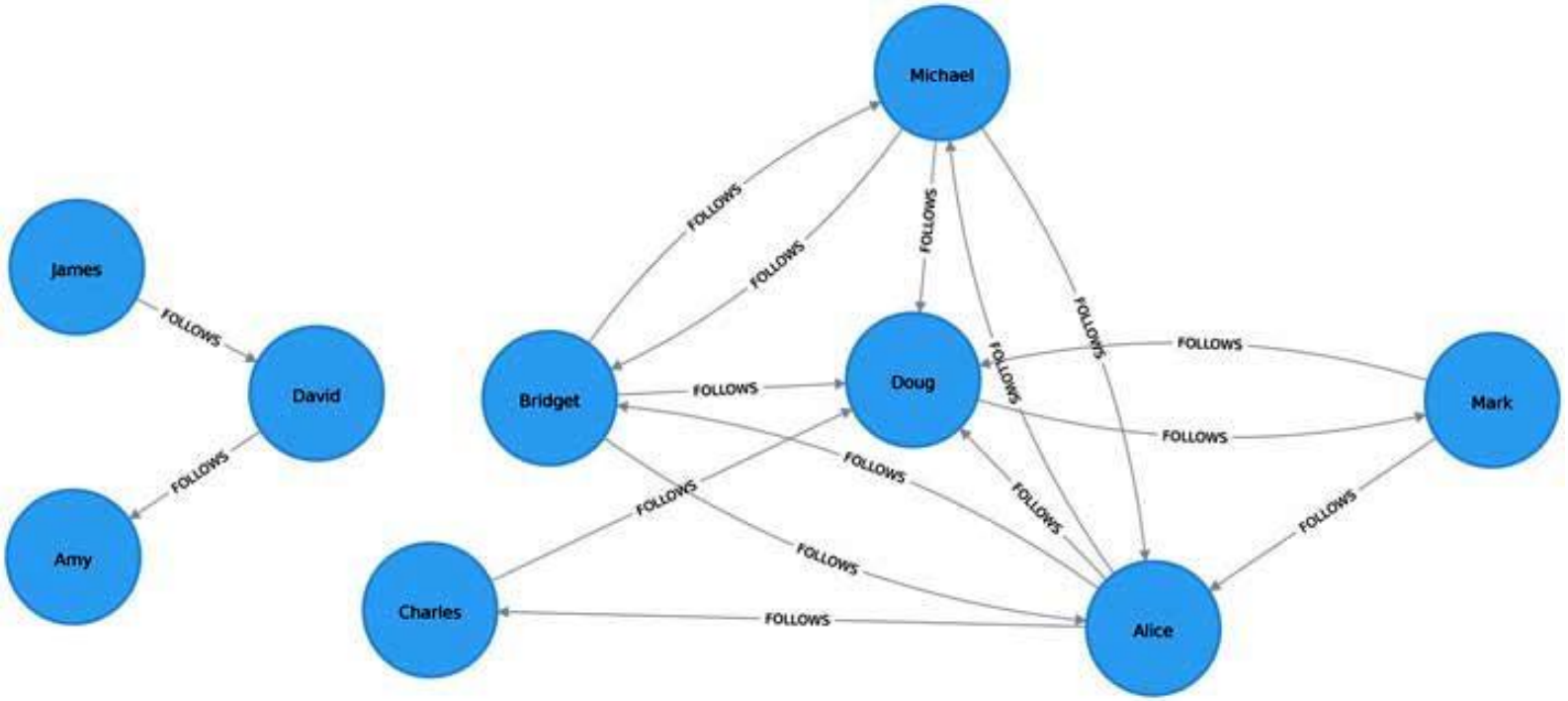
Exemplu de Model Graf Social

src	dst	relationship
Alice	Bridget	FOLLOWS
Alice	Charles	FOLLOWS
Mark	Doug	FOLLOWS
Bridget	Michael	FOLLOWS
Doug	Mark	FOLLOWS
Michael	Alice	FOLLOWS
Alice	Michael	FOLLOWS
Bridget	Alice	FOLLOWS
Michael	Bridget	FOLLOWS
Charles	Doug	FOLLOWS
Bridget	Doug	FOLLOWS
Michael	Doug	FOLLOWS
Alice	Doug	FOLLOWS
Mark	Alice	FOLLOWS
David	Amy	FOLLOWS
James	David	FOLLOWS

Tabel cu relațiile sociale între noduri graf (persoane) (relationships) fisier csv.

id
Alice
Bridget
Charles
Doug
Mark
Michael
David
Amy
James

Tabel cu noduri sociale (persoane-id)(fiser csv)



1. Degree Centrality (Grad de Centralitate)

- Degree Centrality (DC) este cel mai elementar algoritm de acest tip.
  DC **contorizează** numărul de relații de intrare și de ieșire dintr-un nod
  și este folosit pentru **a găsi noduri populare într-un graf**.
- Degree Centrality a fost propus de Linton C. Freeman în lucrarea sa din 1979
  „Centrality in Social Networks: Conceptual Clarification”.

Întinderea (Reach) Înțelegerea întinderii unui nod este o măsură justă a importanței nodului. **La câte alte noduri se poate ajunge din nodul curent?**

- **Gradul unui nod** este numărul de relații directe pe care le are, calculate pentru gradul interior și pentru gradul exterior.

  De exemplu, o persoana cu grad mare într-o rețea de socializare activă ar avea o mulțime de contacte imediate și ar fi mai probabil să se îmbolnăvească de gripă deplasându-se în rețeaua sa.

Gradul mediu al unei rețele este numărul total de relații împărțit la numărul total de noduri; poate fi puternic denaturat de nodurile de grad înalt.

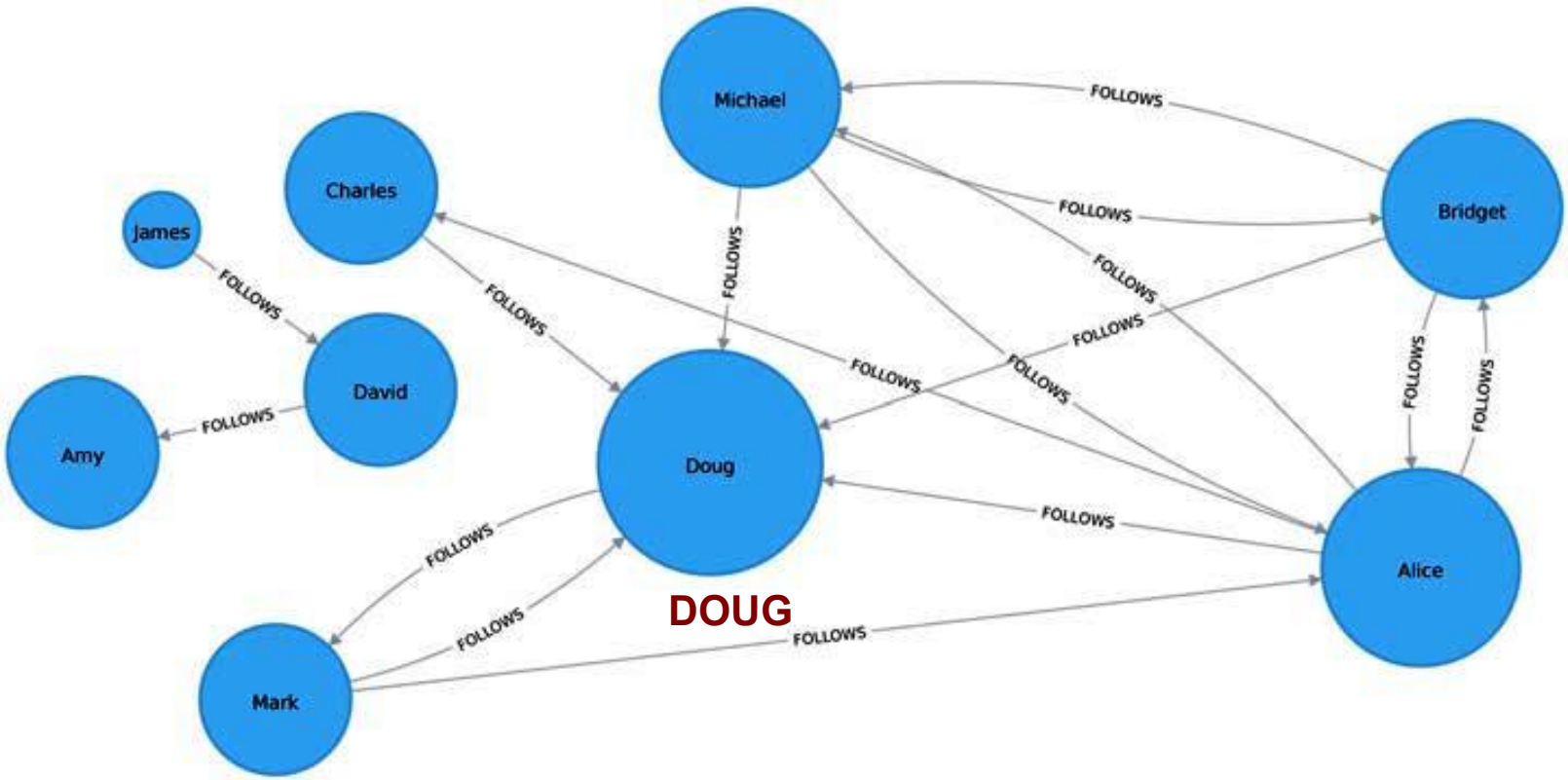
Distribuția gradelor este probabilitatea ca un nod selectat aleator să aibă un anumit număr de relații.

1. Degree Centrality (Grad de Centralitate)

id	degree	inDegree	outDegree
Doug	6	5	1
Alice	7	3	4
Michael	5	2	3
Bridget	5	2	3
Charles	2	1	1
Mark	3	1	2
David	2	1	1
Amy	1	1	0
James	1	0	1

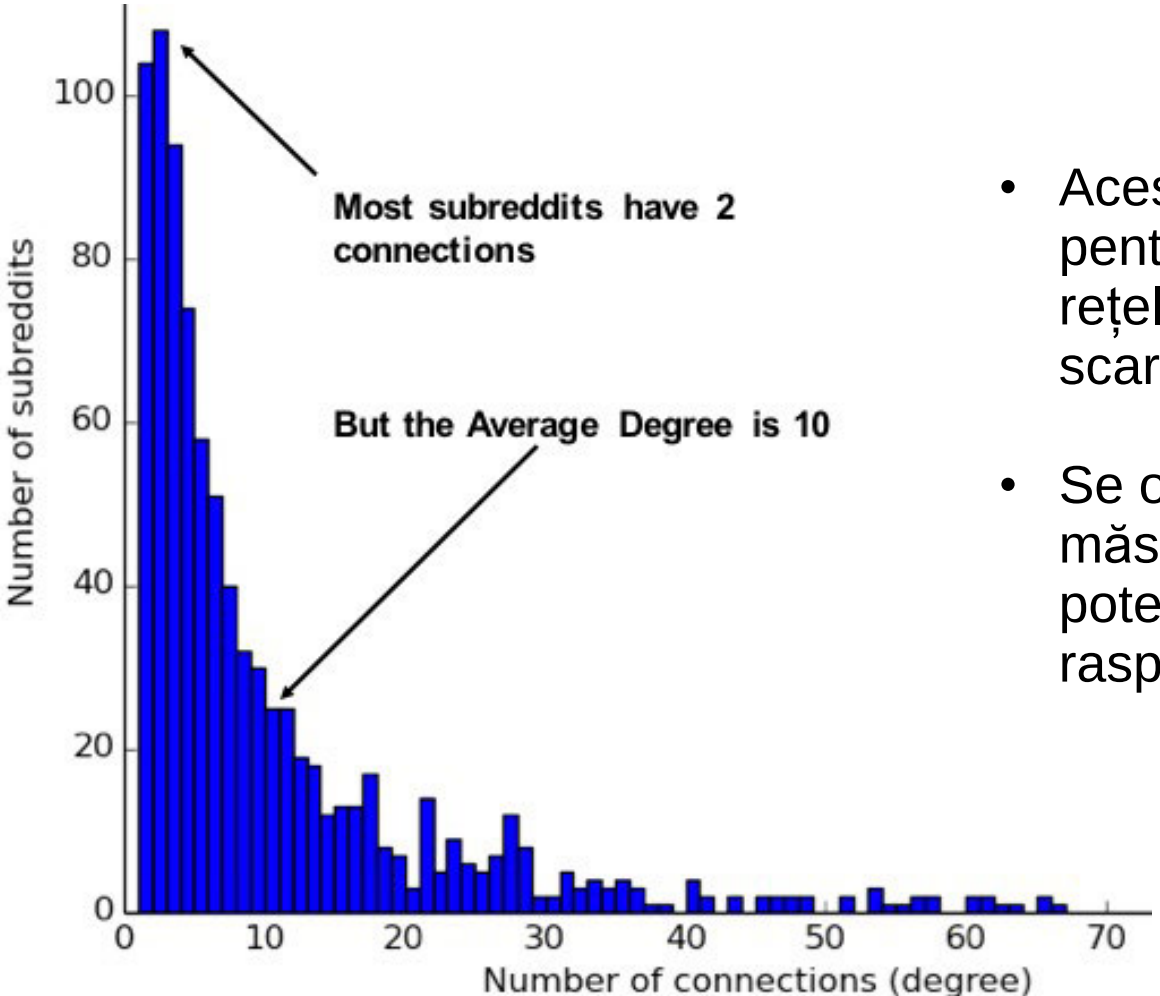
Graful Social cu Centrality Degree

ex. **Doug** este cel mai popular, **in-degree=5 (max.)**



1. Degree Centrality (Grad de Centralitate)

- **Figura ilustrează diferența privind distribuția reala a conexiunilor între subiectele subreddit.**
- Dacă s-ar considera doar media, s-ar presupune că majoritatea subiectelor au 10 conexiuni, în timp ce, de fapt, majoritatea subiectelor au doar 2 conexiuni.



- Aceste măsuri sunt folosite pentru a clasifica tipurile de rețele, cum ar fi rețelele fără scară sau rețelele small-world.
- Se oferă, de asemenea, o măsură rapidă de estimarea potențialului ca lucrurile să se raspândească în rețea.

1. Degree Centrality (Grad de Centralitate)

Când se utilizează Gradul de centralitate?

- Utilizați gradul de centralitate la analiza influenței pe baza numărului de relații de intrare și de ieșire sau determinați „popularitatea” nodurilor individuale.
- Funcționează pentru determinarea conexiunii imediate sau de probabilități pe termen scurt.
- Degree Centrality se aplică și analizei globale pentru evaluarea gradului minim, maxim, gradului mediu și abaterea standard pe întregul graf.

Exemple de cazuri de utilizare includ:

- Identificarea unor indivizi puternici prin relațiile lor, cum ar fi conexiunile oamenilor într-o rețea socială.

  De exemplu “Most Influential Men and Women on Twitter 2017”, de la BrandWatch, primele 5 persoane din fiecare categorie au peste 40 de milioane de urmăritori fiecare.

- Separarea fraudatorilor de utilizatorii legitimi ai unui site de licitații online.

  Centralitatea ponderată a fraudatorilor tinde sa fie semnificativ mai mare din cauza vizării creșterii artificiale a prețurilor.

  Detalii în lucrarea lui P. Bangcharoensap et al. “Two Step Graph-Based Semi-Supervised Learning for Online Auction Fraud Detection”.

2. Closeness Centrality (Apropiere de Centralitate)

- Apropierea de centralitate este un mod de a **detecta nodurile capabile să raspândească informația eficient printr-un subgraf.**
- Măsura centralității unui nod este **distanța medie (inversă) față de toate celelalte noduri.**
- Nodurile cu scor mare de apropiere au cele mai scurte distanțe față de toate celelalte noduri.
- Pentru fiecare nod, algoritmul Closeness Centrality calculeaza suma distanțelor sale față de toate celelalte noduri, pe baza calculului celor mai scurte căi dintre toate perechile de noduri.
- Suma rezultată este apoi inversată pentru a determina scorul de centralitate de apropiere pentru acel nod.

Formula $$C(u) = \frac{1}{\sum_{v=1}^{n-1} d(u,v)}$$

- n este numarul de noduri din graf
- d(u,v) este distanța pe calea cea mai scurtă dintre un alt nod v și u.

Normalizare scor

- Uzual se normalizează acest scor astfel încât să reprezinte **lungimea medie a celor mai scurte căi,** în loc de suma lor.
- Aceasta ajustare permite comparații ale apropierii de centralitate a nodurilor din graf de diferite dimensiuni.

Apropiere de Centralitate normalizată are formula: $$C_{norm}(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(u,v)}$$

2. Closeness Centrality (Apropiere de Centralitate)

Când se utilizează Apropierea de centralitate?

- Se utilizează când trebuie stabilit **ce noduri raspândesc lucrurile cel mai rapid.**
- Utilizarea relațiilor ponderate poate fi foarte utilă în **evaluarea vitezelor de interacțiune în comunicare și analize comportamentale.**

Exemple de cazuri de utilizare (Use Case) includ:

- Descoperirea indivizilor în poziții foarte favorabile pentru a controla și obține informații și resurse vitale în cadrul unei organizații.

- Se utilizează ca euristică pentru estimarea timpului de sosire în telecomunicații și livrarea pachetelor, unde conținutul circulă prin cele mai scurte căi către o țintă predefinită.

- Se poate utiliza pentru clarifica propagarea prin toate căile cele mai scurte simultan, cum ar fi infecțiile ce se raspândesc într-o comunitate locală.

  Detalii în “Centrality and Network Flow”, de S. P. Borgatti.

- Evaluarea importanței cuvintelor dintr-un document, pe baza unui proces de extragere a frazelor cheie.

  Detalii în „A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction”de F. Boudin.

2. Closeness Centrality (Apropiere de Centralitate)

Observație

- Apropierea de Centralitate funcționeaza cel mai bine pe grafuri conectate.
- Când formula originală este aplicata unui graf neconectat, se obține o distanța infinită între două noduri între care nu există drum.
- Deci se ajunge la un scor de centralitate de apropiere **infinit** pentru suma tuturor distanțelor de la acel nod.
- Pentru a evita aceasta problema, se utilizează o variantă a formulei originale.
- *(closeness to others within their subgraph but not the entire graph)*

deci se implementează pe componente/ pe subgrafuri.

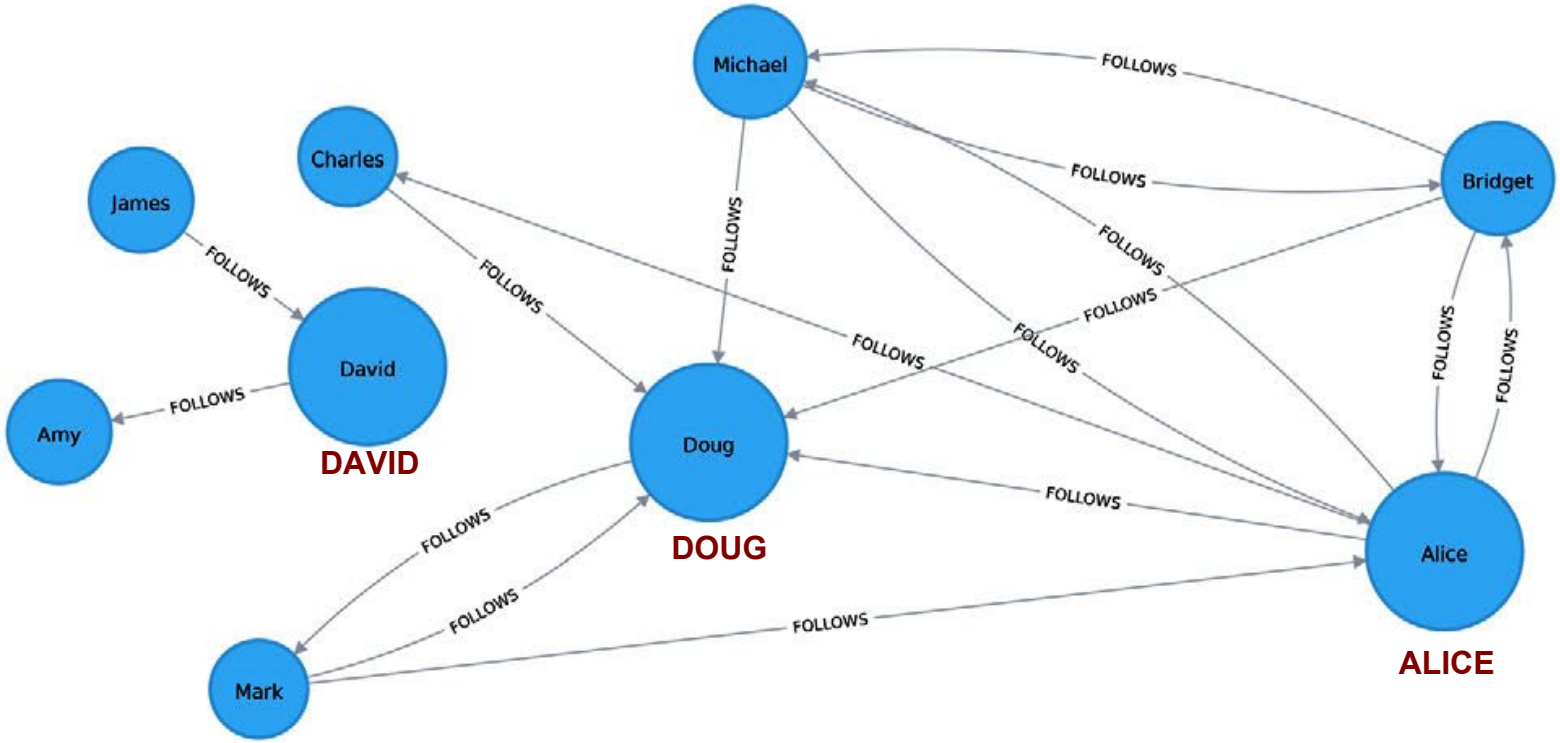
$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(u,v)}$$

- ***u** este* nod.
- ***n*** este nr. Noduri din aceeasi componenta (**subgraf** sau **grup)** ce conține nodul ***u***
- ***d(u,v)*** este calea cea mai scurtă între alte noduri ***v*** și ***u***.

2. Closeness Centrality (Apropiere de Centralitate)

id	ids	closeness
Doug	[[Charles, 1], [Mark, 1], [Alice, 1], [Bridget, 1], [Michael, 1]]	1.0
Alice	[[Charles, 1], [Mark, 1], [Bridget, 1], [Doug, 1], [Michael, 1]]	1.0
David	[[James, 1], [Amy, 1]]	1.0
Bridget	[[Charles, 2], [Mark, 2], [Alice, 1], [Doug, 1], [Michael, 1]]	0.7142857142857143
Michael	[[Charles, 2], [Mark, 2], [Alice, 1], [Doug, 1], [Bridget, 1]]	0.7142857142857143
James	[[Amy, 2], [David, 1]]	0.6666666666666666
Amy	[[James, 2], [David, 1]]	0.6666666666666666
Mark	[[Bridget, 2], [Charles, 2], [Michael, 2], [Doug, 1], [Alice, 1]]	0.625
Charles	[[Bridget, 2], [Mark, 2], [Michael, 2], [Doug, 1], [Alice, 1]]	0.625

- ***Alice, Doug,*** și ***David*** sunt cele mai aproape conectate noduri din graf cu scorul de 1.
- Adică se conectează direct cu toate celelate noduri ale grafului
- ***David,*** desi are **doar două conexiuni (James si Amy)** in graful întreg, dar in grupul de prieteni este important si deci este cel mai apropiat nod de fiecare **i**n subgraf dar nu in întreg graful.



2. Closeness Centrality (Apropiere de Centralitate)
Varianta - **Wasserman și Faust**

- Stanley Wasserman și Katherine Faust au inclus o formula îmbunatățită pentru calcularea apropierii de centralitate pentru grafurile cu mai multe subgrafuri fără conexiuni între acele grupuri.

Detalii în cartea, ***Social Network Analysis: Methods and Applications.***

- Rezultatul acestei formule este un raport dintre fracția de noduri din grup care sunt accesibile și distanța medie de la nodurile accesibile.

Formula este :

$$C_{WF}(u) = \frac{n-1}{N-1}\left(\frac{n-1}{\sum_{v=1}^{n-1} d(u,v)}\right)$$

• ***u*** *este* nod.

• *N este* numărul total de noduri.

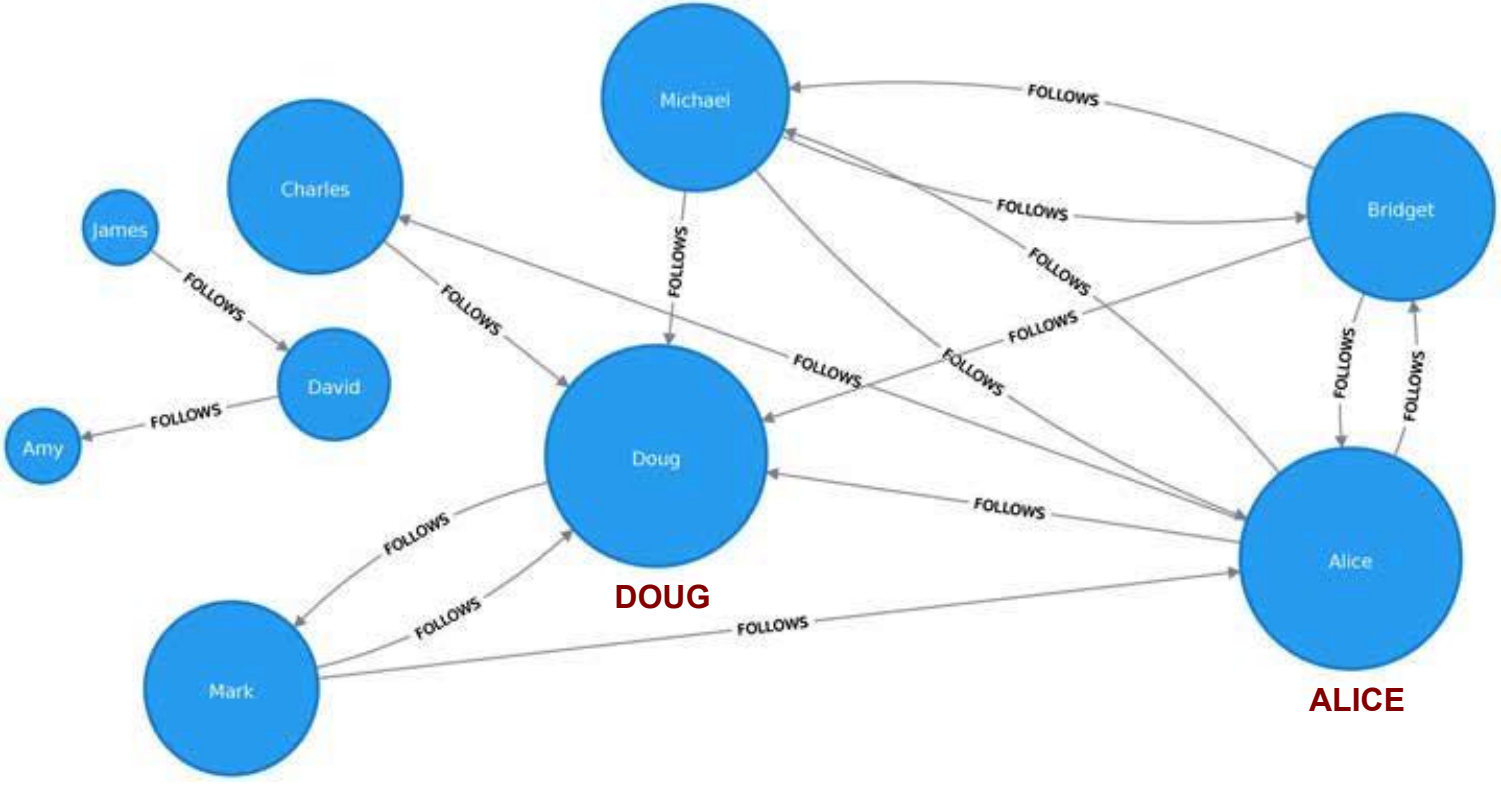
•***n*** este nr. Noduri din aceeasi componenta (**subgraf** sau **grup)** ce conține nodul ***u***

•***d(u,v)*** este calea cea mai scurtă între alte noduri ***v*** și ***u***.

2. Closeness Centrality (Apropiere de Centralitate)
Varianta - **Wasserman și Faust**

- Figura ilustrează rezultatele mai reprezentative de apropierea nodurilor de întregul graf.
  - Scorurile membrilor subgrafului mai mic (**David, Amy și James**) au fost reduse și au acum au **cele mai mici scoruri** dintre toți utilizatorii.
  - Acest lucru este evident deoarece sunt cele mai izolate noduri.
- Aceasta formulă este mai utila pentru a detecta importanța unui nod **pe întregul graf,** decât în cadrul propriului subgraf.
- **Alice si Doug au scorurile cele mai mari de 0.5**

user	centrality
Alice	0.5
Doug	0.5
Bridget	0.35714285714285715
Michael	0.35714285714285715
Charles	0.3125
Mark	0.3125
David	0.125
Amy	0.08333333333333333
James	0.08333333333333333



2. Closeness Centrality (Apropiere de Centralitate)
Varianta - Centralitate armonică

- Centralitatea armonică (Centralitate valoroasă) este o varianta a centralității de apropiere, inventată pentru a rezolva problema originală cu **grafuri neconectate**.

- M. Marchiori și V. Latora (“Harmony in a Small World”) au propus acest concept ca reprezentarea practică a unui drum mediu cel mai scurt (**average shortest path**).

- Calcul scor de apropiere pentru fiecare nod, în loc de suma distanțelor unui nod față de toate celelalte noduri, **se însumeaza inversul acelor distanțe**.

- Acest lucru înseamnă ca **valorile infinite devin irelevante**.

- **Centralitatea armonica** brută pentru un nod este calculată folosind formula:

- n este numarul de noduri din graf
- d(u,v) este distanța cea mai scurta dintre un alt nod v și u.

$$H(u) = \sum_{v=1}^{n-1} \frac{1}{d(u,v)}$$

Centralitatea armonică normalizată are formula:

$$H_{norm}(u) = \frac{\sum_{v=1}^{n-1} \frac{1}{d(u,v)}}{n-1}$$

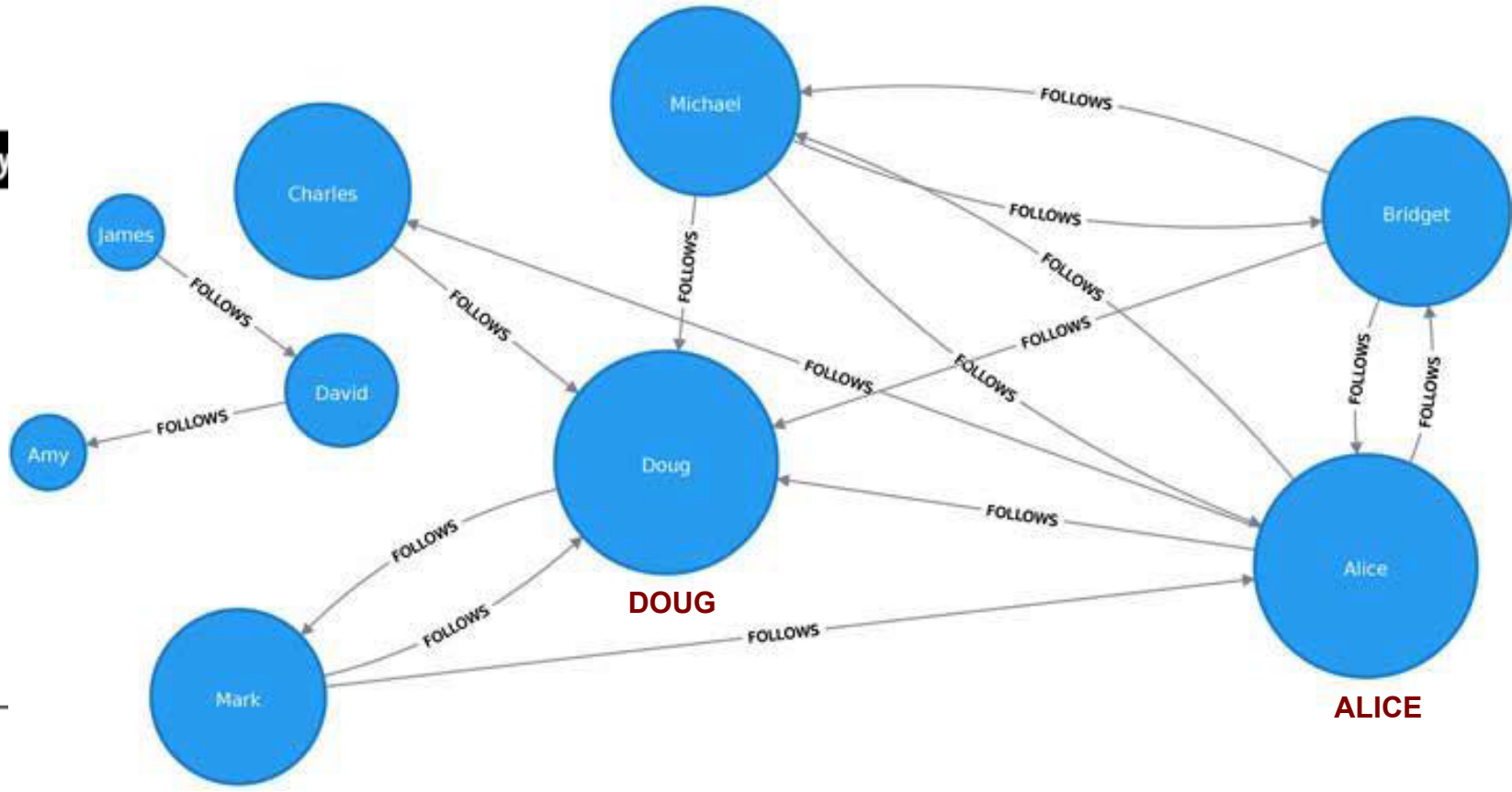
2. Closeness Centrality (Apropiere de Centralitate)
Varianta - **Centralitate armonică**

- Rezultatele diferă de cele ale algoritmului original de apropiere, dar sunt similare cu cele din îmbunatățirea Wasserman și Faust.

  Oricare algoritm poate fi utilizat atunci când lucrați cu grafuri cu mai multe componente conectate.

  Alice si Doug au cele mai mari scoruri de Centralitate armonică similar cu scorurile Waserman șI Faust. (maxime Alice si Doug)

user	centrality
Alice	0.625
Doug	0.625
Bridget	0.5
Michael	0.5
Charles	0.4375
Mark	0.4375
David	0.25
Amy	0.1875
James	0.1875



3) Betweenness Centrality (Inter-Centralitate) (Bridge)

- Uneori, cel mai important nod din sistem nu este cel cu cea mai evidență sau cel mai înalt statut.
- Uneori, intermediarii sunt cei care conectează grupurile sau brokerii sunt cei care controlează cel mai mult resursele sau fluxul de informații.

- Betweenness Centrality este o modalitate de a detecta **cantitatea de influență pe care o are un nod asupra fluxului de informații sau resurse dintr-un graf**.
- Este de obicei folosit pentru a gasi noduri care servesc drept **punte** (**bridge**) dintr-o parte a unui graf la alta.

- **Algoritmul Betweenness Centrality calculeaza mai întâi calea cea mai scurtă (ponderata) între fiecare pereche de noduri dintr-un graf conectat.**
    - Fiecare nod primește un scor, bazat pe numărul acestor cele mai scurte căi care trec prin nod.
    - Cu cât se afla mai multe căi mai scurte pe care se afla un nod, cu atât scorul său este mai mare.

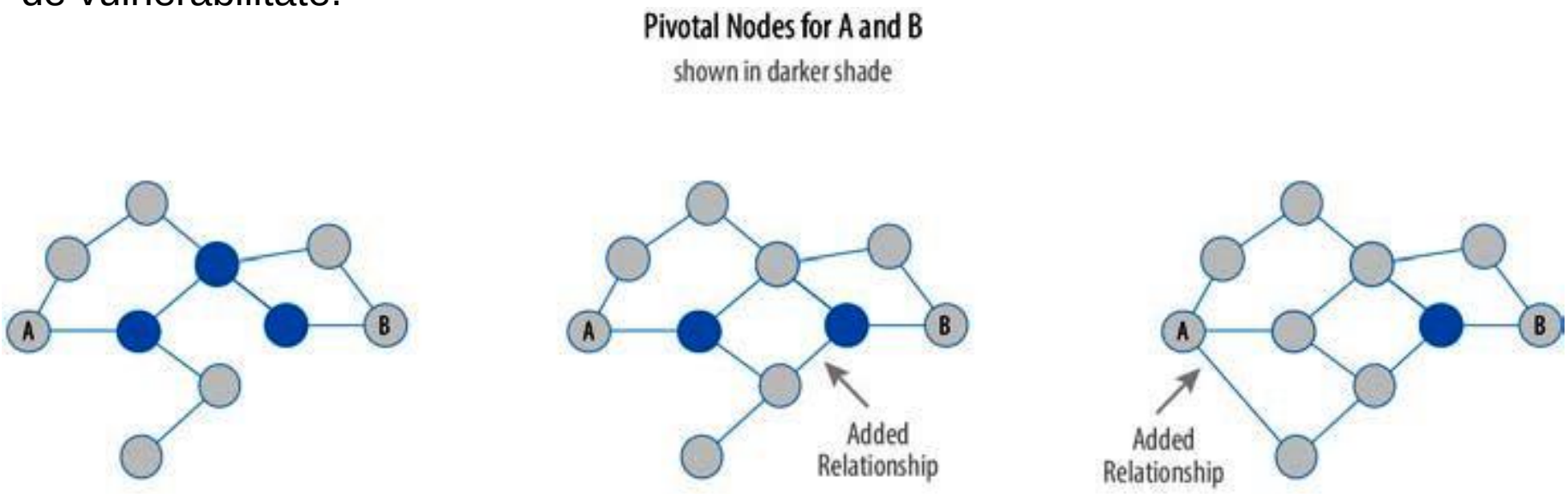
- Betweenness Centrality a fost considerat unul dintre cele „**trei concepte intuitive distincte despre centralitate**” când a fost introdus în lucrarea din 1971, „A Set of Measures of Centrality Based on Betweenness” (L.C. Freeman

3) Betweenness Centrality (Inter-Centralitate) (Bridge)

Punți (Bridges) & Puncte de control

- O punte (bridge) într-o rețea poate fi un nod sau o relație.
- Într-un graf foarte simplu, se găsesc cautând nodul sau relația care, dacă ar fi eliminată, ar determina deconectarea unei secțiuni a grafului.

  Deoarece acest lucru nu este practic într-un graf obișnuit, folosim un algoritm Betweenness Centrality.
- Se poate măsura și distanța dintre un cluster **tratând grupul ca un nod.**
- Un nod este **nod pivot** pentru alte două noduri <u>dacă se află pe fiecare cale cea mai scurtă dintre aceste noduri</u>
- **Nodurile pivot** au rol important în conectarea altor noduri — dacă se elimină un nod pivot, noua cale mai scurtă pentru perechile de noduri originale va fi mai lungă sau mai costisitoare. Aceasta se poate utiliza pentru evaluarea punctelor individuale de vulnerabilitate.



3) Betweenness Centrality (Inter-Centralitate) (Bridge)

Formula de calcul

Se adăugă rezultatele formulei pentru toate căile cele mai scurte,

$$B(u) = \sum_{s \neq u \neq t} \frac{p(u)}{p}$$

unde

- ***u*** este nodul curent
- ***p*** este numărul total al celor mai scurte căi între nodurile ***s*** și ***t.***
- ***p(u***) este numărul celor mai scurte căi dintre nodurile ***s*** și ***t*** care trec prin nodul ***u***.

Procedura:

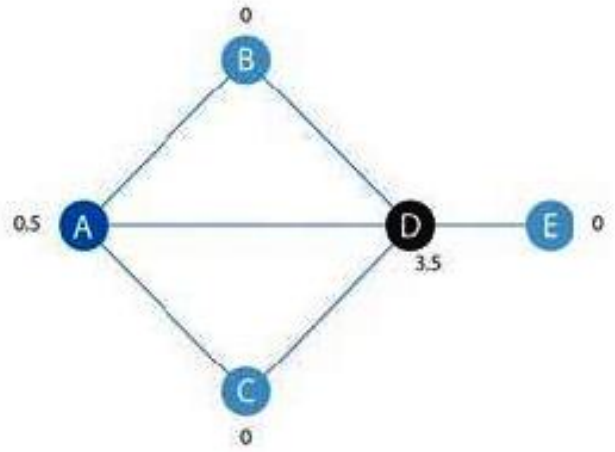
1. Pentru fiecare nod, **se găsesc cele mai scurte căi ce trec prin nod**.

   A. B, C, E nu au cele mai scurte căi și li se atribuie o valoare de 0.

2. Pentru fiecare cale cea mai scurtă din pasul 1, **se calculează procentul acesteia din totalul căilor cele mai scurte posibile pentru acea pereche**.
3. **Se adună toate valorile din pasul 2** pentru a găsi scorul de centralitate al unui nod.

Tabelul din Figură ilustreaza pașii 2 și 3 pentru nodul D.

4. Se repetă procesul pentru fiecare nod.



Node D Calculation

Pairs with Shortest Paths Through D	Total Possible Shortest Paths for That Pair	% of Total Through D (1/Total)
A, E	1	1
B, E	1	1
C, E	1	1
B, C	2 (through D & A)	0.5
Betweenness Score		3.5

3) Betweenness Centrality (Inter-Centralitate) (Bridge)

- **Când se utilizează Betweenness Centrality?**
  Betweenness Centrality se aplică unei game largi de probleme din rețelele din lumea reala. Se utilizează pentru a găsi **blocaje, puncte de control și vulnerabilitați**.
- **Exemple de cazuri de utilizare (Use Case) includ:**
  - **Identificarea influențelor în diverse organizații**.
    Persoanele puternice nu sunt neaparat în poziții de conducere, dar pot fi găsite **în „poziții de intermediere”** folosind Betweenness Centrality.
    Îndepărtarea unor astfel de influenți poate destabiliza serios organizația.
  - Acest lucru ar putea fi o perturbare binevenită de catre organele de aplicare a legii, daca organizația este răufăcătoare, sau ar fi un dezastru dacă o afacere pierde personal cheie pe care l-a subestimat.
    Detalii în “Brokerage Qualifications in Ringing Operations”, by C. Morselli and J. Roy.
- **Descoperirea punctelor cheie de transfer în rețele,** cum ar fi **rețelele electrice**.
  - În mod contraintuitiv, îndepartarea unor punți specifice poate îmbunatăți robustețea generală prin „**insularea”** perturbărilor. („Robustness of the European Power Grids Under Intentional Attack”, de R. Solé, et al.)
- **Ajutând microbloggerii să răspândească acoperirea pe Twitter, cu un motor de recomandare pentru vizarea influențelor**. („Making Recommendations in a Microblog to Improve Impact of a Focal User” de S. Wu et al.)

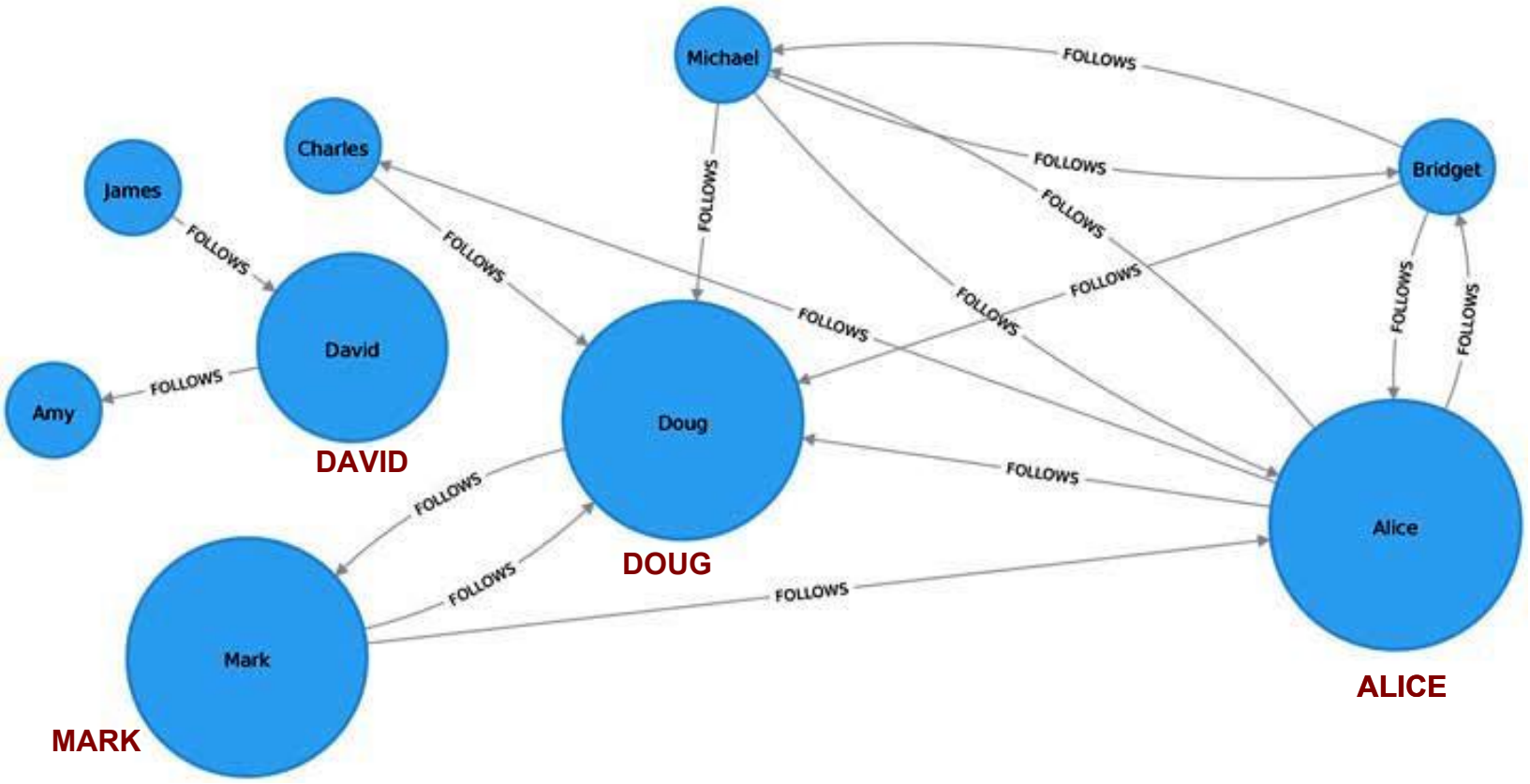
Precizare--Mark Newman în “Networks: An Introduction” Oxford University Press, p186

Betweenness Centrality **presupune că toată comunicarea dintre noduri are loc pe calea cea mai scurtă și cu aceeași frecvență**, ceea ce nu este întotdeauna cazul în viața reală. Deci, nu oferă o vedere perfectă a celor mai influente noduri dintr-un graf, ci doar o bună reprezentare.

3) Betweenness Centrality (Inter-Centralitate) (Bridge)

user	centrality
Alice	10.0
Doug	7.0
Mark	7.0
David	1.0
Bridget	0.0
Charles	0.0
Michael	0.0
Amy	0.0
James	0.0

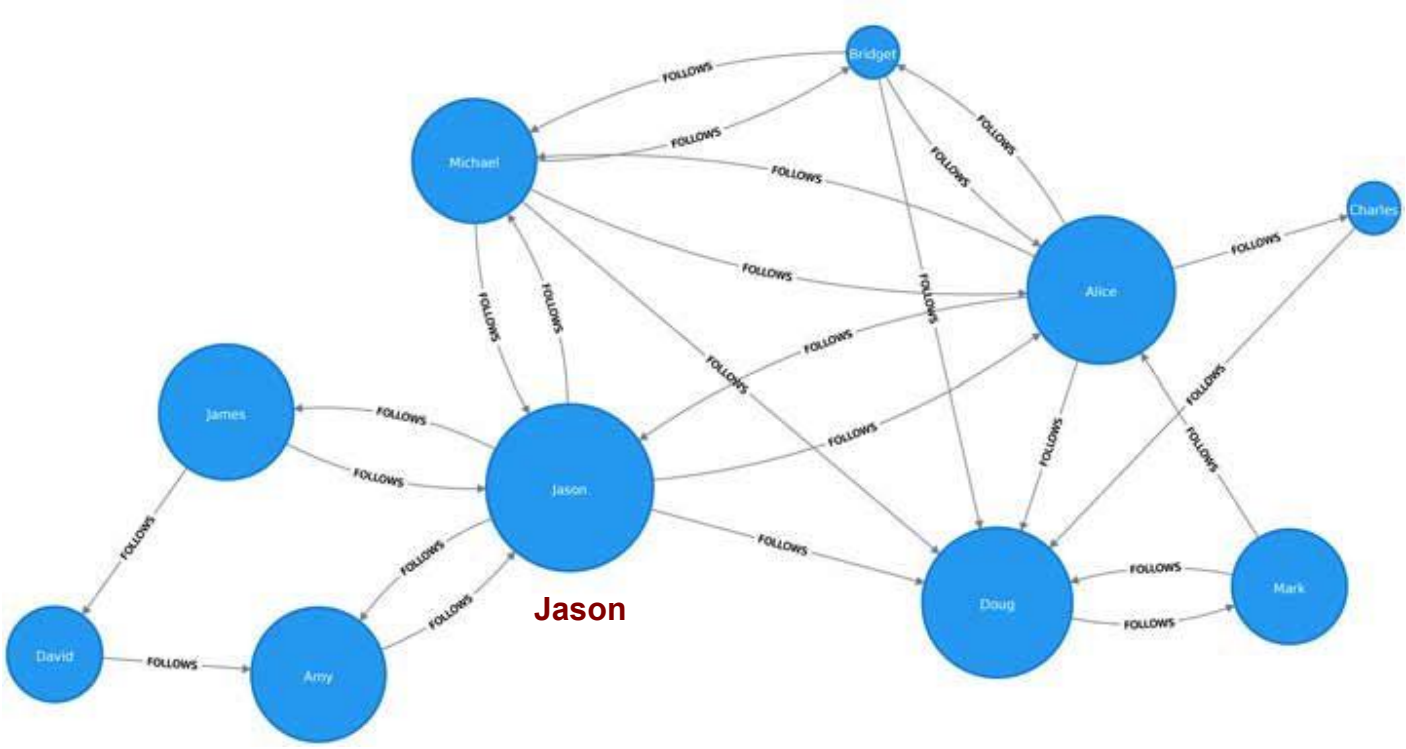
- **Alice este principalul broker** din aceasta rețea
- **Mark și Doug nu sunt departe**.
- În subgraful mai mic, toate căile cele mai scurte trec prin **David**, astfel el este important pentru fluxul de informații între acele noduri.



3) Betweenness Centrality (Inter-Centralitate) (Bridge)

- Pentru grafuri mari, calculul exact al centralității nu este practic.
- **Cel mai rapid** algoritm cunoscut pentru calculul exact între toate nodurile are un **timp de rulare proporțional cu produsul dintre numarul de noduri și numarul de relații.**
- Filtrăm mai întâi la un subgraf sau folosim un subset de noduri.
- Se pot uni cele două componente deconectate prin **introducerea unui nou utilizator** numit **Jason**, ce urmarește (Follows) și este urmat de persoane din ambele grupuri de utilizatori:
- Dacă reluam algoritmul: **Jason are cel mai mare scor,** deoarece comunicarea dintre cele doua seturi de utilizatori va trece prin el.
- ***Jason*** acționeaza ca o punte (**bridge**) locală între cele doua seturi de utilizatori.

user	centrality
Jason	44.33333333333333
Doug	18.333333333333332
Alice	16.666666666666664
Amy	8.0
James	8.0
Michael	4.0
Mark	2.1666666666666665
David	0.5
Bridget	0.0
Charles	0.0



3) Betweenness Centrality (Inter-Centralitate) (Bridge)

Varianta: Randomized-Approximate Brandes (RA-Brandes)

- Calculul exact al **Betweenness Centrality** pentru grafuri mari este foarte costisitor.
- De aceea se poate alege un algoritm de aproximare ce rulează mai rapid, dar oferă totuși informații utile (deși imprecise).
- **Algoritmul RA-Brandes** este cel mai cunoscut algoritm pentru calcularea unui scor aproximativ pentru Betweenness Centrality.
- În loc sa calculeze calea cea mai scurtă între fiecare pereche de noduri, algoritmul RA-Brandes ia în considerare **doar un subset de noduri**.

Doua strategii comune pentru selectarea subsetului de noduri sunt:

1. **Aleator (**Random) Nodurile sunt selectate uniform, la întâmplare, cu probabilitate definita de selecție. **Probabilitatea implicita** este: $\frac{log10(N)}{e^2}$.
   Daca probabilitatea este 1, algoritmul funcționeaza identic cu algoritmul Betweenness Centrality unde sunt încarcate toate nodurile.
2. **Degree -** Nodurile sunt selectate aleator, dar **cele al căror grad este mai mic decât media sunt excluse automat** (adica numai nodurile cu multe relații au șansa de a fi vizitate).

user	centrality
Alice	9.0
Mark	9.0
Doug	4.5
David	2.25
Bridget	0.0
Charles	0.0
Michael	0.0
Amy	0.0
James	0.0

Optimizare Suplimentară - se poate limita adâncimea folosita de algoritmul Shortest path ce va furniza apoi un subset al tuturor căilor cele mai scurte.

Exemplul grafului social – **influencerii de top** sunt similari cu cei de dinainte, (**Alice si Doug**) **deși Mark are un scor mai mare decât Doug**.

Deoarece algoritmul este aleator, vor fi rezultate diferite la fiecare rulare. Eficiența este vizibilă doar la grafuri de dimensiune foarte mare, pentru găsirea unui rezultat rapid, apropiat de cel corect

4) PageRank

- PageRank este cel mai cunoscut dintre algoritmii de centralitate.
    - Măsoară **influența** tranzitivă (sau direcțională) a nodurilor.
    - Toți ceilalți algoritmi de centralitate discutați măsoară influența directă a unui nod, dar **PageRank ia în considerare influența vecinilor unui nod și a vecinilor lor.**
- De exemplu, a avea câțiva prieteni foarte puternici te poate face mai influent decât a avea mulți prieteni mai puțin puternici.
- PageRank este calculat prin distribuirea iterativă a rangului unui nod peste vecinii săi, sau prin parcurgerea aleatorie a grafului și numărând frecvența cu care fiecare nod este parcurs în timpul acestor traversări.
- PageRank este numit după cofondatorul Google, Larry **Page**, care l-a creat pentru a clasifica site-urile web în rezultatele căutarii Google.
- Presupunerea de bază este că o pagină cu mai **multe link-uri de intrare și relații** este mai probabil o **sursa credibilă.**
- PageRank măsoară **numărul și calitatea relațiilor** de intrare într-un nod pentru a determina o estimare a importanței acelui nod.
- Se presupune că nodurile cu influență mai mare asupra unei rețele au mai multe relații de intrare de la alte noduri influente.
- Intuiția din spatele influenței este că relațiile cu nodurile mai importante contribuie mai mult la influența nodului în cauză decât conexiunile echivalente cu nodurile mai puțin importante.

4) PageRank

PageRank este definit in lucrarea originală Google astfel

$$PR(u) = (1 - d) + d\left(\frac{PR(T1)}{C(T1)} + \ldots + \frac{PR(Tn)}{C(Tn)}\right)$$

- Presupunem ca o pagina u are citari de la paginile ***T1*** la ***Tn***.
- ***d*** este un factor de amortizare care este setat între 0 și 1.

Acesta este de obicei setat la 0,85. Poate fi văzut ca probabilitatea ca un utilizator să continue sa acționeze click. Acest lucru ajută la minimizarea scăderii rangului.

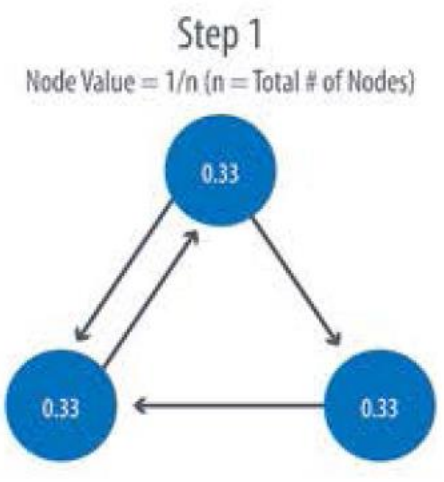
- ***1-d*** este **probabilitatea** ca la un nod să se ajungă direct fără a urma o relație.
- ***C(Tn)*** este gradul exterior al unui nod ***T***.

Observații

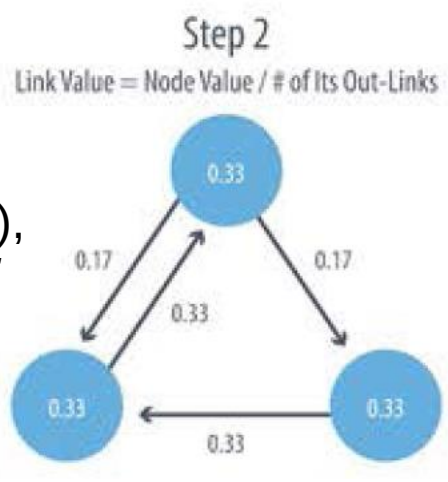
- Măsurarea influenței implică de obicei notarea nodurilor, adesea cu relații ponderate, și apoi actualizarea scorurilor pe mai multe iterații.
- De obicei, toate nodurile sunt utilizate, dar uneori doar o selecție aleatorie este folosită ca distribuție reprezentativă.
- Măsurile de centralitate reprezintă importanța unui nod în comparație cu alte noduri.
- Centralitatea este **o clasare a impactului potențial al nodurilor**, nu o măsură a impactului real.
- De exemplu, se pot identifica 2 persoane cu cea mai mare centralitate într-o rețea, dar poate că sunt considerate diverse politici sau norme culturale și acestea transferă influența asupra altor persoane.
- Cuantificarea impactului real este un domeniu activ de cercetare pentru a dezvolta metrici de influență suplimentare.

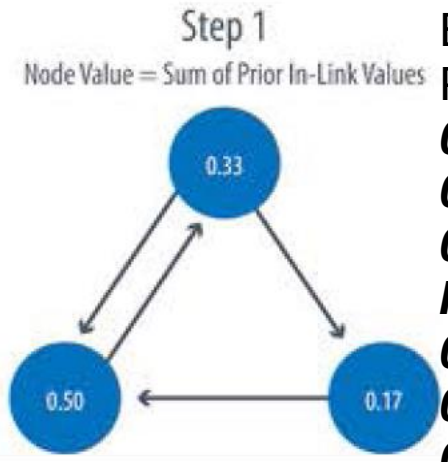
4) PageRank

Figura prezintă un exemplu elementar despre modul în care **PageRank va continua să actualizeze rangul unui nod pâna când acesta converge** sau **îndeplinește numărul stabilit de iterații.**

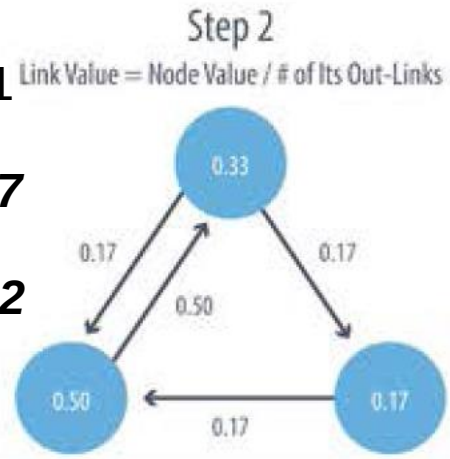


Exemplu
Pass 0 (sus),
0.33/2=0.17
0.33/1=0.33
0.33/1=0.33



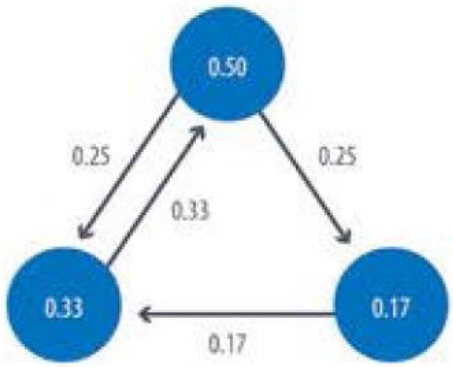


Exemplu
Pass 1-Step 1
0.33=0.33
0.5=0.33+0.17
0.17=0.17
Pass 1-Step 2
0.33/2=0.17
0.50/1=0.50
0.17/1=0.17



Exemplu
Pass 0 (sus),
0.33/2=0.17
0.33/1=0.33
0.33/1=0.33

Pass 1,
Pass 2 (jos)
d=0.85

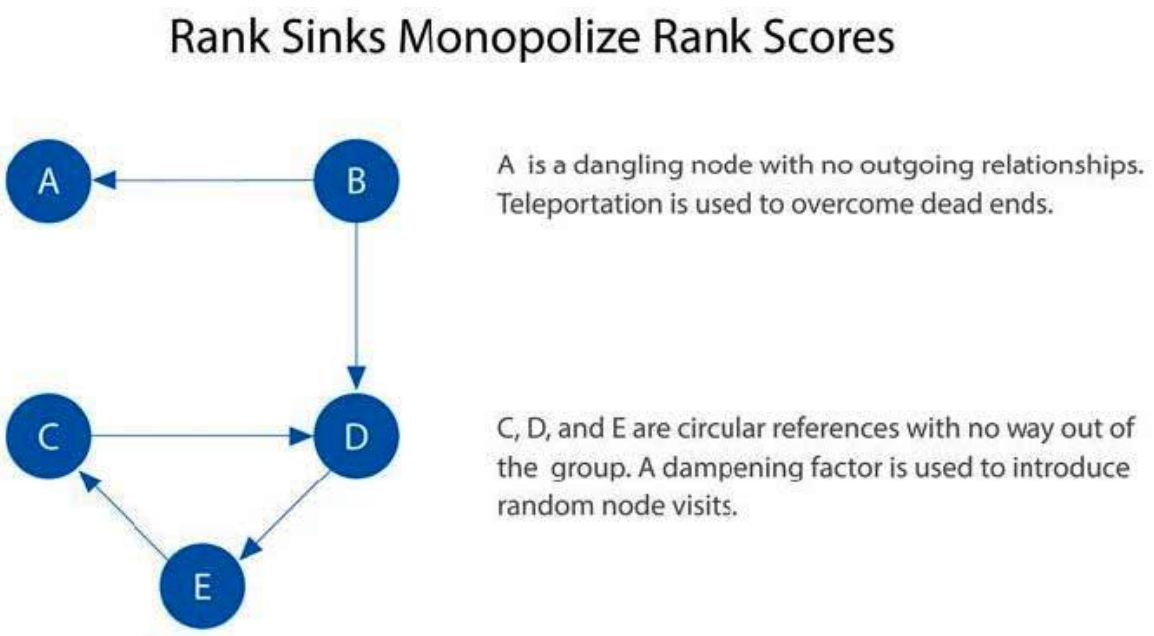


PASS 2-Step 1
0.5=0.5 (in)
0.33=0.17+0.17
0.33=0.33
PASS 2 Step 2
0.5/2=0.25
0.33/1=0.33
0.17/1=0.17

4) PageRank - Iteration, Random Surfers & Rank Sinks

- PageRank este un algoritm iterativ ce ruleaza fie pâna când scorurile converg, fie pâna când se atinge un numar stabilit de iterații.
- Conceptual, PageRank presupune ca există un navigator care vizitează paginile urmând link-uri sau utilizând o adresa URL aleatorie.
- Factorul de amortizare ***d*** definește **probabilitatea ca următorul clic sa fie printr-un link**.
- Se vede ca probabilitatea ca un web-surfer să se plictisească și să treacă aleator la o alta pagină.
- Un scor PageRank reprezintă probabilitatea ca o pagina sa fie vizitată printr-un link de intrare ce nu este aleatoriu.
- Un nod sau un grup de noduri, fără relații de ieșire (numit și nod suspendat) poate monopoliza scorul PageRank. Aceasta este numit ***rank sink.***
- Este similar cu un web-surfer blocat pe o pagina sau pe un subset de pagini, fără nicio ieșire. (ex. **nod A,** fara ieșire)
- Altă dificultate este creată de nodurile ce indică doar unul către celălalt într-un grup.
- **Referințele circulare**

(ex. Figura noduri **C,D,E**) provoacă creșterea rangurilor lor, pe măsura ce web-surferul sare înainte și înapoi printre noduri.



4) PageRank

Exista două strategii folosite pentru a evita scăderile de rang.

1) În primul rând, când se ajunge la un **nod care nu are relații de ieșire**, PageRank presupune relații de ieșire cu toate nodurile.

Traversarea legăturilor invizibile este uneori numita **teleportare.**

2) În al doilea rând, **factorul de amortizare** oferă o altă oportunitate de a evita problemele prin introducerea unei probabilități pentru o legatură directă versus vizitarea aleatorie a nodurilor.

- Când setați ***d=0,85***, un nod complet aleator este vizitat în 15% din timp.
- Deși formula originală recomandă un factor de amortizare de 0,85, utilizarea sa inițiala a fost pe World Wide Web cu o distribuție conform legii de putere a link-urilor (majoritatea paginilor au foarte puține link-uri și câteva pagini au multe).
- Scaderea factorului de amortizare scade probabilitatea de a urma trasee lungi de relație înainte de a face un salt la întâmplare.
- La rândul sau, acest lucru crește contribuția predecesorilor imediați ai unui nod la scorul și rangul sau.
- Daca apar rezultate neașteptate ale PageRank, să realizați o analiză exploratorie a grafului pentru a vedea daca vreuna dintre aceste probleme este cauza.

Detalii în „“The Google PageRank Algorithm and How It Works” de Ian Rogers.

4) PageRank

Când se utilizează PageRank?

- PageRank este acum utilizat în multe domenii în afara indexării web.
- Utilizați acest algoritm ori de câte ori căutați o influență mare asupra unei rețele.
- De exemplu, in biologie, daca vizați o genă cu cel mai mare impact asupra unei funcții biologice, este posibil să nu fie cea mai conectată poate fi, de fapt, gena cu cele mai multe relații cu alte funcții mai semnificative.

Exemple de cazuri de utilizare (Use case) ale PageRank:

- Prezentarea utilizatorilor cu **recomandări pentru alte conturi** pe care ar putea dori sa le urmeze (Twitter folosește PageRank personalizat pentru aceasta).
  - Algoritmul este rulat pe un graf ce conține interese comune și conexiuni comune.

Detalii în lucrarea „The Who to Follow Service at Twitter”, de P. Gupta et al.

- **Predicția fluxului de trafic** și a circulației umane în spațiile publice sau străzi.
  - Algoritmul este rulat pe un graf al intersecțiilor rutiere, unde scorul PageRank reflectă tendința oamenilor de a parca sau termina călătoria pe fiecare stradă.

Detalii în “Self-Organized Natural Roads for Predicting Traffic Flow: A Sensitivity Study”, de B. Jiang et al.

- În **sisteme de detectare a anomaliilor și fraudelor din industria de sănătate și asigurari,** PageRank ajută la dezvăluirea medicilor sau furnizorilor ce se comportă neobișnuit, iar scorurile sunt apoi introduse într-un algoritm de învățare automată.

David Gleich descrie alte utilizări ale algoritmului în „PageRank Beyond the Web”.

4) PageRank

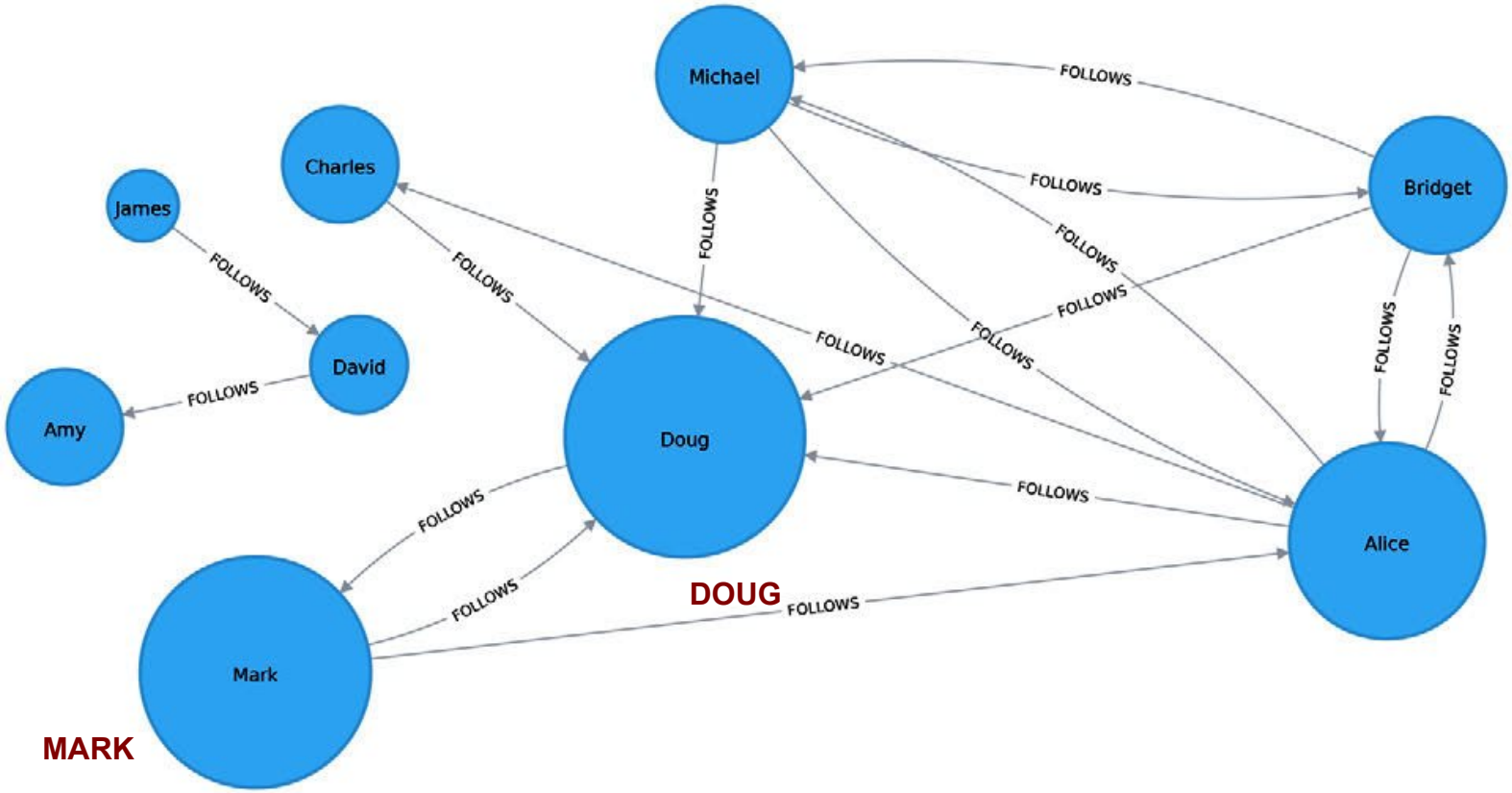
page	score
Doug	1.6704119999999998
Mark	1.5610085
Alice	1.1106700000000003
Bridget	0.535373
Michael	0.535373
Amy	0.385875
Charles	0.3844895
David	0.2775
James	0.15000000000000002

- Exemplu Figura - Rezultat PageRank – ***Doug*** este the most influential user, urmat indeaproape de ***Mark***; se vede importanța nodurilor unul relativ la celalalt.

- Implementarile PageRank diferă, astfel încât pot produce scoruri diferite chiar și atunci când ordinea este aceeași.

Exemplu. inițializează nodurile cu valoarea: 1 minus factor amortizare,

În acest caz, clasamentele relative (obiectivul PageRank) sunt identice, dar valorile scorului de bază utilizate pentru a atinge rezultatele sunt diferite.



4) PageRank personalizat (Varianta PageRank)

- **Personalized PageRank (PPR)** este o varianta a algoritmului PageRank care calculeaza importanța nodurilor într-un graf din perspectiva unui anumit nod.
- Pentru PPR, salturile aleatorii se referă la un set dat de noduri de pornire.
- Astfel se obțin rezultate controversate (biases rezults) sau personalizate pe baza nodului de start.
- Pe baza acestei controverse (bias) și localizare, PPR devine util pentru recomandări foarte bine direcționate.

id	pageRank
Alice	0.1650183746272782
Michael	0.048842467744891996
Bridget	0.048842467744891996
Charles	0.03497796119878669
David	0.0
James	0.0
Amy	0.0

Ex. Graf social rezultate calcul PageRank personalizat, pe cine să urmeze Doug, excluzand pe cei ce sunt deja urmăriți de Doug și pe el insăși.
Rezultat

- **Alice** este cea mai bună sugestie de urmat (to Follow)
- Se pot sugera și următoarele 2 opțiuni cu scoruri egale, Michael și Bridget.

Concluzie

- Algoritmii de centralitate sunt un instrument excelent pentru identificarea influențelor într-o rețea.
- Algoritmii prototip pentru centralitate sunt:
  1) Degree Centrality
  2) Closeness Centrality
  3) Betweenness Centrality și
  4) PageRank.
- Există și variante de rezolvare a problemelor cu durate lungi de execuție și componente izolate, precum și opțiuni pentru utilizări alternative.
- Există multiple utilizări pe scară largă pentru algoritmii de centralitate și se încurajează explorarea acestora pentru analize diverse.
- Se pot aplica cele prezentate pentru a localiza nodurile de contact optime pentru diseminarea informațiilor, găsirea nodurilor ascunse ce controlează fluxul de resurse și descoperirea nodurilor indirecte cu putere in rețea.