Véletlen gráfok és valós hálózatok

Csima Judit

BME, SZIT

2010. február 17.

Tartalom

1. Motiváció: miért pont véletlen gráfok?

2. A klasszikus modell: Erdős-Rényi véletlen-gráf modell
   - definíció
   - jellemzői (legnagyobb komponens mérete, összefüggőség, fokszámeloszlás)

3. Milyen jellemzőik vannak a valós hálózatoknak?

4. Jó (jónak tűnő) modellek a valós hálózatokra

   (a) néhány inhomogén véletlen-gráf modell
   (b) Barabási-Albert féle preferential attachment modell

Motiváció

Valós hálózatok vizsgálata fontos: web, kommunikációs hálózatok, szociális hálózatok, stb.

Nehezen vizsgálhatók: túl nagyok, teljes leírás lehetetlen, nincs látható struktúra, globális viselkedést nem lehet jól leírni

$\Rightarrow$

lokális vizsgálat: milyen szabály szerint, hogyan kapcsolódnak a csúcsok egymáshoz?

$\Rightarrow$Véletlen gráfok, mert a lokális szabályok nem lehetnek determinisztikusak

Kérdés: milyen véletlen-gráf modellt használjunk?

Klasszikus modell: Erdős-Rényi-modell

P. Erdős-A. Rényi: On the evolution of random graph, *Magyar Tud. Akad. Mat. Kutató Int. Közl.,***5**: 17-61, (1960)

statikus, homogén modell

$ER_p(n)$:

- n csúcs
- az egyes élek p valószínűséggel vannak jelen a gráfban (egymástól függetlenül)
  p gyakran $p = \lambda/n$

(E. N. Gilbert: Random graphs, *Ann. Math. Statist.*, **30**: 1141-1144 (1959)

Kitérő: Erdős és Rényi eredeti modellje

$ER(n, M)$: n csúcsú, M élű véletlen gráfok közül egy, egyenletes eloszlás szerint választva

Eredmények eredetileg $ER(n, M)$-re, de innen:

$$P_p(E) = \sum_{M=0}^{n(n-1)/2} P_M(E) \cdot \binom{\binom{n}{2}}{M} p^M (1-p)^{\binom{n}{2}-M}$$

mivel az élszám eloszlása $ER_p(n)$-ben $\binom{n}{2}$ és p paraméterű binomiális eloszlás.

Lényeg: A két modellben kapott eredmények megfeleltethetők egymásnak, mostantól az $ER_p(n)$-nál maradunk.

p megválasztása

$ER_p(n)$ élszámának várható értéke $pn(n-1)/2$

Ritka gráfokat akarunk kapni, ezért természetes választás $p = \lambda/n$, ekkor az élszám várhatóan $\lambda(n-1)/2 = \Theta(n)$

További lehetőség p-re: $p(n)$ függvényként, ahol $p(n) \cdot n$ tart c-hez. (Ennek spec. esete a $p = \lambda/n$, ekkor $p(n) \cdot n \equiv \lambda$, azaz $c = \lambda$.)

Kérdések: mi történik a gráfban λ értékétől (az általános $p(n)$ függvénytől, ill. c-től) függően?

Válaszok: mindig olyasmi, hogy adott paraméter-érték mellett nagy valószínűséggel igaz vmi, azaz egy adott esemény valószínűsége tart 1-hez, ha $n \to \infty$.

Fázisátmenet $p = \lambda/n$ eset

Leghíresebb eredmény az ER modellel kapcsolatban: a maximális komponens méretéről szól

- $\lambda < 1$: sok $\Theta(\log n)$ méretű komponens
- $\lambda = 1$: a max. komponens csúcsszáma $\Theta(n^{2/3})$
- $\lambda > 1$: egy $\Theta(n)$ méretű és sok $\Theta(\log n)$ méretű komponens

Megjegyzés: már kicsi n-ekre is nagy valószínűséggel megtörténik az óriás-komponens kialakulása

A bizonyítás elve

Vegyünk egy $ER_p(n)$ véletlen gráfot és abban egy v csúcsot, becsüljük meg, hogy mekkora lesz az a komponens, amiben v benne van

Jelölés: $C(v) =$ a v csúcs komponense

Kellene $|C_{max}| = \max_{v \in \{1,2,\ldots,n\}} |C(v)|$

Nézzük egyesével ezeket a $C(v)$-ket, kezdjük pl. az 1 csúcs komponensével

Képzeletben járjuk be ezt szélességi bejárással:

X_1= az 1 távolságra levő csúcsok száma, azaz 1 közvetlen szomszédainak száma

Ez $n-1$ és p paraméterű binomiális eloszlású valószínűségi változó, azaz $P(X_1 = k) = \binom{n-1}{k} p^k (1-p)^{(n-1-k)}$

Bizonyításvázlat/2

Jelölés: $X_1 \sim BIN(n-1, p)$

Jelölje az 1 csúcs közvetlen szomszédait $i_1 < i_2 < \cdots < i_{X_1}$ és nézzük most i_1 új szomszédait, ezeknek száma legyen X_2.

X_1 értékét rögzítve, i_1 új szomszédainak száma, X_2, eloszlása ismét binomiális, de most a paraméterek $n-1-X_1$ és p, azaz $X_2 \sim BIN(n-1-X_1, p)$

Általában: ha $C(1)$ $i+1$. csúcsának az új szomszédait nézzük, akkor ezek számának feltételes eloszlása a korábbi X_i-ket adottnak véve $X_{i+1} \sim BIN(n-1-X_1-X_2-\ldots-X_i, p)$

Bizonyításvázlat/3

Mikor áll le a komponens bejárása?

Ha már nincs olyan csúcs $C(1)$-ben, aminek a szomszédait ne vizsgáltuk volna.

Az $i.$ csúcs vizsgálata után a még nem vizsgált $C(1)$-beli csúcsok száma $1 + X_1 + X_2 + \cdots + X_i - i$, azaz ez az eljárás leáll, ha

$1 + X_1 + X_2 + \cdots + X_i - i = 0.$

Tehát $|C(1)| = min\{i \,:\, X_1 + X_2 + \cdots + X_i = i - 1\}$

Bizonyításvázlat/4

Keressük tehát

$|C(1)| = min\{i \,:\, X_1 + X_2 + \cdots + X_i = i - 1\}$-et.

Baj: X_i-k nehezen kezelhetők, nem is függetlenek

Nade: nagy n-re $BIN(n, \lambda/n)$ közel van a λ paraméterű Poisson-eloszláshoz, azaz

$$P(BIN(n, \lambda/n) = k) \approx e^{-\lambda}\frac{\lambda^k}{k!}$$

Ha tehát a binomiális eloszlású X_i-k helyett független, azonos, Poisson-eloszlású ${X_i}^*$-okra gondolunk, akkor

$$|C(1)| = min\{i \,:\, {X_1}^* + {X_2}^* + \cdots + {X_i}^* = i - 1\}$$

Bizonyításvázlat/5

Kellene $|C(1)| = min\{i : 1 + {X_1}^* + {X_2}^* + \cdots + {X_i}^* = i\}$, ahol ${X_i}^*$-ok független, azonos, Poisson-eloszlásúak.

Ez viszont olyan, mint egy branching process, Poisson-eloszlás-számnyi leszármazottal:

kezdetben adott egy csúcs (1), gyerekeinek száma $({X_1}^*)$ λ-paraméterű Poisson-eloszlást követ, majd minden leszármazott csúcsra ugyanez igaz.

$1 + {X_1}^* + {X_2}^* + \cdots + {X_i}^* = i$ azt jelenti, hogy a populáció első i tagjának összes leszármazottjainak (beleértve az őst is) halmaza megegyezik a populáció első i tagjának halmazával, azaz a populáció az i. tagnál kihal.

Mikor történik ez?

Ez sokat tanulmányozott dolog, itt ismert jelenség vmi fázisátmenet-szerű dolog:

A gyerekek számának várható értékétől, λ-tól függően:

- $\lambda \leq 1$: 1 vgel kihal a populáció
- $\lambda > 1$: pozitív vgel örökké él

Innen ered az $ER_{\lambda/n}(n)$-modell fázisátmenetének magyarázata:

- $\lambda < 1$: sok $\Theta(\log n)$ méretű komponens
- $\lambda = 1$: a max. komponens csúcsszáma $\Theta(n^{2/3})$
- $\lambda > 1$: egy $\Theta(n)$ méretű és sok $\Theta(\log n)$ méretű komponens

Az ER modell további tulajdonságai

1. Fázisátmenet, ha $p(n)$ nem feltétlenül λ/n alakú, de $p(n) \cdot n$ tart vmi c konstanshoz:

- $c < 1$: sok $\Theta(\log n)$ méretű komponens
- $c > 1$: egy $\Theta(n)$ méretű és sok $\Theta(\log n)$ méretű komponens

$c = 1$ esetben részben nyitott: a max. komponens mérete a $p(n) \cdot n - 1$ nagyságrendjén múlik, ismert, hogy mi van, ha ez a különbség nagyobb, mint $\log n \cdot n^{-1/3}$

Ekkor a max. komponens mérete pl. lehet $\Theta(n^{2/3} \log n)$ is.

Az ER modell további tulajdonságai/2

2. Mikor lesz összefüggő egy $ER_{\lambda/n}(n)$ gráf?

Fix λ esetén nagy valószínűséggel nem összefüggő.

- $\lambda(n) - \log n \to \infty$ esetén nagy valószínűséggel összefüggő.
- $\lambda(n) - \log n \to -\infty$ esetén nagy valószínűséggel nem összefüggő.

3. Fokszámeloszlás

Mindegyik csúcs fokszámeloszlása ugyanolyan , $BIN(n-1, \lambda/n)$, azaz határeloszlásban λ paraméterű Poisson; a tipikus fokszám megegyezik a várható értékkel, kicsi a szórás.

Tehát: ez a modell nem lesz jó valós hálózatok leírására.

Milyenek a valós hálózatok?

Az világos, hogy nem olyanok, mint amilyenek az ER-modell gráfjai.

Sokáig nem sokat lehetett tudni róluk, mert túl nagyok, de ahogy nőtt a számolási kapacitás egyre több dolog derült ki.

Lényeg: sok közös vonás van a különböző helyről származó hálózatokban (www, telekommunikációs hálózatok, neurális hálózatok, szociális hálózatok, stb.).

Legfontosabb jellemzők:

1. Small world: kicsi az átlagos távolság a csúcspárok között

(Ez sokszor magyarázható a hálózat jellegéből, de nem mindig.)

Milyenek a valós hálózatok?

2. Skála-független fokszámeloszlás: a k-ad fokú csúcsok aránya $k^{-\tau}$-val arányos, azaz hatványeloszlás szerint csökkenik (persze $\tau > 1$)

Azért hívják skála-függetlennek, mert ez a jelenség már kevés csúcs esetén is jelentkezik (azaz most ”scale-free”= hatványeloszlás szerinti fokszámok)

Ábrázolás gyakran log-log skálán, mert ha $N_k =$ a k-ad fokú csúcsok arányára igaz, hogy

$N_k = c \cdot k^{-\tau}$, akkor

$\log N_k = \log c - \tau \cdot \log k,$

azaz az így kapott függvény lineáris és meredeksége éppen $-\tau$.

Kérdések: 1. τ mekkora? (2. És honnan jön ez a fokszámeloszlás?)

Példák valós hálózatokra

1. Internet (maga a vezetékhálózat): csúcsok az Autonomous System-ek, élek az ilyenek közti közvetlen vezetékek

1. átlagos távolság (pl. hány AS-en megy keresztül egy email) kicsi (≈ 7) $\rightarrow$ small-world

2. power-law fokszámeloszlás, $\tau \approx 2.2$

2. Szociális hálózatok (”six degrees of separation”)

Milgram leveles kísérlete (1967): 6 a max. távolság két ember között az ismeretségi gráfban

Ugyanez a kísérlet email-lel 2001-ben: átlagos út a célig 6 hosszú $\rightarrow$ small-world

Az ismeretségi gráf fokszámeloszlása: nehezen vizsgálható a teljes gráf, ehelyett email-gráf (ki kinek küld levelet) ill. iwiw, Facebook, stb. vizsgálata.

Az látszik, hogy skála-független fokszámeloszlás, τ kérdéses, általában $\approx 1.8 - 2.5$

3. Színészes gráfok

Melyik színész melyik másik színészekkel játszott közös filmben?

átlagos távolság 6 két színész között

fokszámeloszlás majdnem power-law: egy darabig $k^{-\tau}$, de a vége exponenciálisan cseng le, $\tau \approx 2.3$

Hasonlók igazak a társszerzős gráfokra is.

4. WWW

Csúcsok a weboldalak, élek a linkek

$\tau_{in} \approx 2.1$, $\tau_{out} \approx 2.45$,
(bár van olyan vélemény is, hogy itt nincs is power-law, az csak mérési hibaként jön ki)

átlagos távolság (irányítatlan eset)≈ 7

A valós hálózatokra tehát igaz, hogy ..

1. Kicsi az átlagos távolság
2. Ritkák: élszám $\Theta(n)$
3. Fokszámaik hatványeloszlás szerint

Tehát: olyan modell kell, ami ezeket a jellemzőket (főleg a hatványeloszlást) képes

(a) reprodukálni $\rightarrow$ statikus modell: fix csúcsszám

(b) generálni, magyarázni $\rightarrow$ dinamikus modell: növekvő csúcsszám

Először most a statikus modelleket nézzük meg.

Általánosított véletlen-gráf modell

Az ER modell homogén gráfokat eredményezett, mert minden egyes élnek uakkora volt a megjelenési valószínűsége

Ehelyett: az élek egymástól függetlenül vannak jelen, de a valószínűségük a végpontjaikhoz rendelt súlyoktól függ,

a súlyok lehetnek állandóak vagy lehetnek maguk is valváltozók

Például: az ij él valószínűsége legyen $p_{ij} = \frac{w_i w_j}{l_n + w_i w_j}$, ahol $l_n = \sum_{i=1}^{n} w_i \rightarrow$ $GRG_n(\mathbf{w})$ gráf, ha w_i fixek, illetve

az ij él valószínűsége legyen $p_{ij} = \frac{W_i W_j}{\mu n + W_i W_j}$, ahol W_i-k független, azonos eloszlású valószínűségi változók μ várható értékkel $\rightarrow$ $GRG_n(\mathbf{W})$ gráf

Általánosított véletlen-gráf modell

Ha $w_i \equiv \frac{n\lambda}{n-\lambda}$, akkor $p_{ij} = \lambda/n \ \rightarrow$ visszakaptuk az eredeti ER modellt

Ha a W_i független, azonos eloszlású valószínűségi változók hatványeloszlásúak a keletkező véletlen gráf fokszámeloszlása hatványeloszlás lesz

Mivel a gráf definíciója lényegében tartalmazza a hatványeloszlást, nem nagy szám, hogy ez kijött...

Konfigurációs modell

Szintén statikus modell, (lényegében) tetszőleges rögzített fokszámeloszláshoz gyárt egy (nagyjából) megfelelő gráfot.

A fokszámeloszlás itt is lehet fix vagy lehet valószínűségi változó által meghatározott.

Minden csúcshoz hozzárendeljük a kívánatos fokszámot $\rightarrow$ képzeletben ráillesztünk fokszámnyi fél-élet, majd a fél-éleket párosítjuk

Ez persze nem feltétlenül lesz egyszerű gráf (pedig mi olyat szeretnénk) $\rightarrow$

Kérdés: hogyan csináljunk egyszerű gráfokat?

Konfigurációs modellek

1. módszer: ha fix fokszámok vannak: töröljük a hurokéleket és a többszörös élekből csak egy maradjon $\rightarrow$ törlős konfigurációs modell

Belátható, hogy a törlés nem nagyon változtatja meg a fokszámokat, a kapott gráf lényegében jó lesz.

2. módszer: ha a fokszámok független, azonos eloszlású valváltozókból jönnek:

belátható, hogy ha a fokszámok nem túl nagyok, akkor pozitív valószínűséggel egyszerű gráfot kapunk $\rightarrow$ ismételjük addig az eljárást, amíg egyszerű gráf jön ki $\rightarrow$ ismétléses konfigurációs modell

Ha azt írjuk elő, hogy a fokszámok hatványeloszlás szerint legyenek, akkor az így kapott véletlen gráf jó lesz, tudja a skála-függetlenséget.

Dinamikus modell

De ez megint olyan volt, hogy azért jött ki a power-law, mert esélye se volt nem-kijönni. Nem magyarázza, hogy a valós hálózatokban miért jelenik meg a power-law mindig.

Az első olyan modell, ami magyaráz is: Barabási-Albert féle preferential attachment modell

- csúcsonként növeszti a hálózatot leíró gráfot
- a motiváló heurisztikáról hihető, hogy a valós helyzetet írja le
- nem nyilvánvaló módon vezet power-law fokszámeloszláshoz

Fizikusok definíciója, matematikusok szerint pontatlan, de precízzé tehető

Barabási-Albert-féle preferential attachment modell

Egyre nagyobb csúcsszámú gráfot hoz létre

Motiváció: az új csúcsok nagyobb valószínűséggel kapcsolódnak a nagyobb fokszámú régiekhez

Ez hihető pl. a szociális hálózatoknál (valószínű, hogy egy új ember ismerősei inkább a szociálisan aktívak közül kerülnek ki)

a színészes gráfban (egy kezdő színész valószínűleg sokat játszó színészekkel fog először szerepelni filmben)

www-nél: egy új oldal valószínűleg az ismertebb weboldalakra tartalmaz linket

Barabási-Albert-féle eredeti definíció

Kezdjünk néhány (m_0) izolált csúccsal

minden lépésben egy új csúcsot veszünk be m éllel úgy, hogy ezek az élek mind régi csúcsokhoz vezetnek

annak a valószínűsége, hogy az új csúcs egy régi x_i csúcsot választ egy élének végpontjául legyen $d(x_i)/\sum_j d(x_j)$.

Így tehát t lépés múlva lesz $m_0 + t$ csúcs és mt él.

Barabási-Albert: szimulációkkal kijön, hogy a fokszámeloszlás hatványeloszlás $\tau = 3$-mal.

A definíció pontosítása

Az eredeti definíció több ponton nem is jó ill. nem elég precíz:

1. Mivel izolált pontokkal kezd, nem világos, hogy hogyan lesznek egyáltalán élek, hiszen egy új él behúzásának valószínűsége $d(x_i)/\sum_j d(x_j)$, azaz $\equiv 0$.

Megoldás: valami rögzített m_0 csúcsú, e_0 élű gráffal kezdünk

2. Ha egy új él valószínűsége $d(x_i)/\sum_j d(x_j)$, akkor az új csúccsal együtt behúzott élek számának várható értéke nem m, hanem 1.

Megoldás: 1. az új él valószínűsége $m(d(x_i)/\sum_j d(x_j))$

2. definiáljuk a modellt $m = 1$-re, majd ennek segítségével tetszőleges m-re.

A definíció pontosítása/2

3. Nem világos, hogy hogyan veszi be az m új élet az új csúccsal együtt:

Egyszerre?

A baj ekkor az, hogy az m új él eloszlása nincs egyértelműen meghatározva a $m(d(x_i)/\sum_j d(x_j))$ valószínűségek által:

ha pl. egy 4-hosszú körhöz veszünk hozzá $m = 2$-vel egy új csúcsot,

akkor tetszőleges olyan ”egyszerre két élet”-választás jó, ahol két szomszédos csúcsot $0 \leq \alpha \leq 1/4$, két nem-szomszédos csúcsot pedig $1/2 - 2\alpha$ valószínűséggel választunk,

mert ilyenkor mindegyik csúcs valóban $m \cdot (d(x_i)/\sum_j d(x_j)) = 2 \cdot 2/8 = 1/2$

valószínűséggel választódik.

A definíció pontosítása/3

Vagy az élek egymás után, egyesével kerülnek be?

De ilyenkor melyik fokszám számít, az új csúcs megjelenése előtti vagy a már néhány új éllel növelt?

4. Lehetnek-e többszörös élek, hurokélek?

Lényeg: minden probléma megoldható, a definíció precízzé tehető és

lényegében bárhogyan tesszük precízzé, igazolható a fokszámokra a hatványeloszlás.

Egy lehetséges precízzé tevés

Kezdetben 1 csúcs, egy hurokéllel

$m = 1$-re a növekedés: az t. új csúcsnak, v_t-nek, egy új éle van, ennek másik végpontja $\frac{1}{2t-1}$ valószínűséggel saját maga, $\frac{d(v_j)}{2t-1}$ valószínűséggel egy régebbi v_j csúcs

$m \geq 2$ esetben: mt csúcsszámú gráfot hozunk létre a fenti módon, mintha $m = 1$-es gráfot csinálnánk, egyesével bevéve az csúcsokat, 1-1 éllel,

aztán m-esével összehúzzuk a csúcsokat

$\rightarrow$ lesz t csúcs és mt él, ahogy kell

Ez olyasmi, mintha egyesével vennénk be az új éleket és közben minden él után frissítenénk a fokszámot.

Eredmények

Számos szimulációban kijön a hatványeloszlás

Akárhogy tesszük precízzé a definíciót belátható a hatványeloszlás

További paramétereket lehet bevezetni a modellben, lehet még általánosítani, így bármely $3 \geq \tau > 2$ kitevő kihozható

$m = 1$ esetén a gráf kevés, nagy-méretű komponensből áll, a legnagyobb nagyságrendje $\Theta(\frac{t}{\log t})$
$m \geq 2$ esetén nagy valószínűséggel összefüggő lesz a kapott gráf

$m = 1$ esetén a gráf átmérője nagy valószínűséggel $\Theta(\log t)$
$m \geq 2$ esetén a gráf átmérője nagy valószínűséggel $\Theta(\frac{\log t}{\log \log t})$

Irodalom

1. Remco van der Hofstad: Random Graphs and Complex Networks, *lectures notes for master courses*

http://www.win.tue.nl/ rhofstad/NotesRGCN2008.pdf

2. A.-L.-Barabási, R. Albert: Emergence of scaling in random networks. *Science*, **286**(5439), 509-512, (1999)

3. B. Bollobás, O. Riordan, J. Spencer, G. Tusnády: The degree sequence of a scale-free random graph process. *Random Structures Algorithms*,**18**(3), 279-290, (2001)

5. B. Bollobás, O. Riordan: Random graphs and branching processes, *Handbook of Large-Scale Random Networks*, Bolyai Society Math. Studies, 18, Chapter 1 (2009?)