



Stockholms
universitet

En jämförelse av triangelskapande mekanismer hos slumpgrafer

Tobbe Anliot

Kandidatuppsats 2015:16
Matematisk statistik
Juni 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

En jämförelse av triangelskapande mekanismer hos slumpgrafer

Tobbe Anliot*

Juni 2015

Sammanfattning

I denna uppsats studerar vi två olika slumpmekanismer som kan användas för att öka transitiviteten i Erdős-Rényi klassiska slumpgraf $G(n, p)$. Verkliga sociala nätverk uppvisar vanligen en hög grad av transitivitet, dvs. dina vänner är ofta också vänner sinsemellan. I en slumpgraf representeras detta genom många trianglar, något som saknas i $G(n, p)$.

Den ena metoden går ut på att alla öppna tripletter, dvs. 3 noder med två kanter som sammanbinder dem, stängs genom att med någon sannolikhet addera den tredje kanten och skapa en triangel. Den andra metoden bygger på att man skapar en ny triangelgraf där var och en av alla möjliga trianglarna utplaceras med någon sannolikhet, och sedan skapas unionen av triangelgrafen och $G(n, p)$ där eventuella dubbelkanter tas bort.

I uppsatsen bestämmer vi slutna formler för det förväntade antalet kanter och trianglar och studerar också gradfördelningen. Vi undersöker också den globala klusterkoefficienten som anger andelen stängda tripletter. Vi bestämmer den förväntade klusterkoefficienten analytiskt för den senare metoden och visar varför den är svår att bestämma för den tidigare. Vi undersöker också grafernas största komponent, där två noder tillhör samma komponent om det finns en väg av kanter som sammanbinder dem. Vi visar med simuleringar att om vi håller det förväntade antalet kanter och trianglar fixt så skiljer sig den största komponentens storlek åt mellan de två metoderna.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: tobbe.anliot@gmail.com. Handledare: Pieter Trapman, Maria Deijfen, Cecilia Holmgren.

Tack

Jag vill passa på och tacka mina handledare Pieter Trapman, Maria Deijfen och Cecilia Holmgren för att förtjänstfullt och med mycket tålamod introducerat mig till området slumpgrafer.

Innehåll

1	Introduktion	4
2	Erdős-Rényi graf $G(n, p)$	4
2.1	Grad- och kantfördelning hos $G(n, p)$	5
2.2	Förväntat antal trianglar	5
2.3	Förväntad klusterkoefficient	6
2.4	Triangelutökning enligt metod 1	7
2.4.1	Förväntat antal kanter och trianglar	7
2.4.2	Gradfördelning	10
2.5	Triangelutökning enligt metod 2	11
2.5.1	Förväntat antal kanter och trianglar	11
2.5.2	Gradfördelning	13
3	En jämförelse av den största komponenten hos de två metoderna	13
4	En jämförelse av klusterkoefficienten hos de två metoderna	15
4.1	Förväntad klusterkoefficient med metod 2	15
4.2	Förväntad klusterkoefficient med metod 1	16
5	Diskussion	17

1 Introduktion

I den här uppsatsen kommer vi att studera olika triangelskapande mekanismer hos slumpgrafer. Den enklaste och mest studerade modellen är Erdős-Rényi grafen $G(n, p)$ där varje möjlig kant mellan två noder adderas med en sannolikhet p oberoende av övriga kanter och n är antalet noder i grafen.

I många verkliga grafer, ex. i sociala nätverk, hittar vi en hög grad av transitivitet, det vill säga dina vänner är i stor utsträckning också vänner. I en graf med hög grad av transitivitet hittar vi alltså många trianglar och detta är inte något man ser hos $G(n, p)$ varför den passar illa som en modell för verkliga sociala nätverk. Mot den bakgrunden är det därför intressant att utforska två olika slumpmekanismer som syftar till att öka antalet trianglar till en befintlig graf $G(n, p)$.

1. Grafen $G(n, p)$ innehåller en mängd öppna tripletter. En öppen tripplett består av tre noder sammanbundna av två kanter, se figur 1. För varje sådan öppen tripplett adderar vi nu med någon sannolikhet den sista kanten och skapar en triangel. Denna metod kommer vi härnäst att kalla metod 1.
2. I en graf $G(n, p)$ finns $\binom{n}{3}$ möjliga trianglar. Var och en av dessa trianglar adderas till $G(n, p)$ med någon sannolikhet. Eventuella dubbelkanter som uppstår tas slutligen bort. Denna metod finns översiktligt beskriven i [3] och kommer härnäst kallas för metod 2.

Slumpgrafers beteende studeras vanligen i fallet där antalet noder n går mot oändligheten och så också i detta arbete. Vi kommer att bestämma det förväntade antalet kanter och trianglar för grafer skapade enligt de två metoderna och också försöka säga något om gradfördelningen, dvs. antalet grannar en nod har. Ett vanligt mått på transitiviteten i en graf är dess globala klusterkoefficient och vi kommer att visa varför det är svårt att bestämma dess förväntade värde i den först föreslagna metoden ovan. En intressant egenskap hos $G(n, p)$ är storleken hos dess största komponent [1, s. xiv], där två noder tillhör samma komponent om det finns en väg av kanter som sammanbinder dem. Vi kommer därför med hjälp av simuleringar jämföra storleken på denna komponent, hos de två metoderna.

2 Erdős-Rényi graf $G(n, p)$

En slumpgraf består av en uppsättning noder $V = (v_1, \dots, v_n)$ och kanter, $E = (e_1, \dots, e_m)$ som sammanbinder par av noder (v_i, v_j) . De noder som har en kant till nod v kallas grannar till v och antalet grannar en nod har brukar benämnas nodens grad, d . Två noder säger vi tillhör samma komponent om det finns en väg av kanter som sammanbinder dem.

Detta arbete tar sin utgångspunkt i den slumpgraf som Erdős och Rényi började studera på sent 50-tal, i den fortsatta texten benämnd ER-graf eller $G(n, p)$. Modellen $G(n, p)$ anger nu att alla $\binom{n}{2}$ möjliga kanter mellan de n noderna existerar med sannolikheten p , oberoende av de övriga.

2.1 Grad- och kantfördelning hos $G(n, p)$

Graden d hos en slumpvist vald nod i $G(n, p)$ är binomialfördelad med parametrar $(n - 1)$ och p . Den förväntade graden är därför $(n - 1)p$ och vi ser alltså att det förväntade antalet grannar till en nod går mot oändligheten då $n \rightarrow \infty$. I verkliga sociala nätverk har vi inte ett oändligt antal vänner och då vill efterlikna verkliga nätverk sätter vi att $p = \frac{\lambda}{n-1}$ och fördelningen kan då approximeras med en poissonfördelning, se [2, s. 171], med parameter λ dvs. förväntat antal grannar λ är oberoende av n .

Antalet kanter i grafen är också binomialfördelat men med parametrar $\binom{n}{2}$ och $\frac{\lambda}{n-1}$, då vi har $\binom{n}{2}$ kanter att välja bland och alla inkluderas oberoende med sannolikheten $\frac{\lambda}{n-1}$. Återigen kan denna binomialfördelning approximeras med en poisson fördelning, $Po(\binom{n}{2} \frac{\lambda}{n-1}) = Po(\frac{n\lambda}{2})$ med förväntat antal kanter $\frac{n\lambda}{2}$.

Vad har vi då för beroendeförhållande mellan graderna hos de olika noderna då n är stort?

Definition 1. Om $D_{v_i}^{(n)}$ är gradtalet för nod v_i i en slumpgraf med n noder säger vi att gradtalen för två noder v_i och v_j är asymptotiskt oberoende om

$$P(D_{v_i}^{(n)} = k | D_{v_j}^{(n)} = l) \rightarrow P(D_{v_i}^{(n)} = k), \text{ då } n \rightarrow \infty.$$

I ER-grafen har vi att om $l = 0$ så är $D_{v_i}^{(n)} \sim \text{Bin}(n - 2, \frac{\lambda}{n-1})$ -fördelad vilken kan approximeras med $Po(\lambda)$. Om $l > 0$ så är sannolikheten att $e_{i,j}$ är en av dessa l kanter $1 - \frac{n-2}{n-1} \cdot \frac{n-3}{n-2} \cdot \dots \cdot \frac{n-l-1}{n-l} = \frac{l}{n-1}$, vilken går mot noll då n går mot oändligheten. Detta betyder att med en sannolikhet som går mot 1 när n går mot oändligheten så är återigen $D_{v_i}^{(n)} \sim \text{Bin}(n - 2, \frac{\lambda}{n-1})$ -fördelad och kan alltså approximeras med en $Po(\lambda)$ -fördelning för stora n . Vi har alltså att nodernas gradtal alla är asymptotiskt parvist oberoende. För stora n är de också approximativt oberoende.

2.2 Förväntat antal trianglar

Innan vi beräknar det förväntade antalet trianglar går vi igenom några begrepp och definitioner. En tripplett är en mängd bestående av tre olika noder. Om dessa är sammanbundna av exakt två kanter säger vi att vi har en öppen tripplett centrerad i den mittersta noden och om de är sammanbundna av tre kanter kallas den stängd, centrerad i någon av de tre möjliga noderna, se figur 1. Varje triangel kan då representeras av tre stängda trippletter centrerade i varsitt hörn av triangeln.

Figur 1: Öppen tripplett(vänster) och stängd tripplett(höger).



Hittills har vi inte talat om ordnade trippletter men vi kommer att använda dessa i senare avsnitt varför vi diskuterar dem redan nu. En ordnad tripplett är en lista på de tre noderna som ingår där ordningen är av betydelse. Om vi har en stängd tripplett, $[v_1, v_2, v_3]$, centrerad i v_2 så betyder det att vi har 2 ordnade stängda trippletter centrerade i v_2 , nämligen $[v_1, v_2, v_3]$ och $[v_3, v_2, v_1]$. En konsekvens av denna definition är bland annat att en triangel representeras av 6 ordnade stängda trippletter.

Med definitionerna ovan kan vi nu bestämma de förväntade antalet trianglar i en graf $G(n, p)$. Om vi låter T vara mängden av alla möjliga trippletter av noder, och därmed alla möjliga trianglar så har vi att T består av $t = \binom{n}{3}$ element. Låt nu $X_1, \dots, X_t$ vara stokastiska indikatorvariabler där $X_i = 1$ om den i :te mängden av tre noder är en triangel och $X_i = 0$ annars. Då alla X_i är oberoende och likafördelade bernoullivariabler med parameter $Be(p^3)$ ger linjariteten hos väntevärdet att

$$E\left[\sum_{i=1}^t X_i\right] = \sum_{i=1}^t E[X_i] = tp^3 = \frac{t\lambda^3}{(n-1)^3} = \frac{n(n-1)(n-2)\lambda^3}{6(n-1)^3} \approx \frac{\lambda^3}{6}$$

då $n \rightarrow \infty$. Vi kan alltså konstatera att det förväntade antalet trianglar kommer att vara litet i relation till n då $n \rightarrow \infty$ vilket alltså motiverar att vi utökar grafen med ytterligare trianglar vilket är vad denna uppsats handlar om.

2.3 Förväntad klusterkoefficient

I verkliga nätverk ser vi ofta grupper, eller kluster, av tätt sammanbundna individer. Sannolikheten att individer i en sådan grupp känner varandra är betydligt större än den genomsnittliga sannolikheten att två slumpvist valda individer i nätverket känner varandra. Den globala klusterkoefficienten C är ett mått på till vilken grad noder samlas i kluster och definieras

$$C = \frac{3 \cdot \text{antalet trianglar}}{\text{antalet sammanbundna trippletter}} = \frac{\text{antalet stängda trippletter}}{\text{antalet sammanbundna trippletter}}.$$

där antalet sammanbundna trippletter definieras som summan av antalet öppna och stängda trippletter, se [6, s. 12].

Vi kan börja med att konstatera att i en ER-graf har vi visat att det förväntade antalet trianglar är begränsat, oberoende av n , approximativt $\frac{\lambda^3}{6}$, då n är stort. Vi vet också att antalet grannar d till en given nod är $Po(\lambda)$ -fördelat, oberoende av n , vilket medför $P(D > 1) > 0$, då $n \rightarrow \infty$. Detta medför i sin tur att sannolikheten att en given nod v är i centrum av en öppen tripplett är större än 0 då $n \rightarrow \infty$. Vi kan därför sluta oss till att klusterkoefficienten $C \rightarrow 0$, då $n \rightarrow \infty$.

Vi kan också hitta en sluten formel för klusterkoefficienten. Förväntat antal trianglar är alltså approximativt $\frac{\lambda^3}{6}$. För att tre noder ska vara sammanbundna, alltså en sammanbunden tripplett, krävs att det existerar två kanter mellan dem. Förväntat antal sammanbundna trippletter är därför $\binom{n}{3}(\frac{\lambda}{n-1})^2 \approx \frac{\lambda^2(n-1)}{6}$. Vi får följaktligen att den förväntade klusterkoefficienten $E[C] \approx \frac{\lambda^3}{6} / \frac{\lambda^2(n-1)}{6} = \frac{\lambda}{n-1}$ är nära 0 för stora n . Detta resonemang bygger dock på att antalet trianglar och antalet sammanbundna trippletter är asymptotiskt oberoende. Detta kan dock inses vara fallet om man betänker att antalet trianglar är en konstant och antalet sammanbundna trippletter går mot oändligheten, då n går mot oändligheten.

2.4 Triangelutökning enligt metod 1

Att utöka grafen med ett antal trianglar enligt metod 1 går nu ut på att identifiera alla trippletter sammanbundna av två kanter och med någon sannolikhet addera den tredje kanten för att skapa en triangel. För att senare kunna jämföra de två metoderna vill vi kunna hålla det förväntade antalet trianglar och kanter hos de två metoderna jämbördiga. Därför börjar vi nu med att undersöka vad det förväntade antalet kanter och trianglar är om en graf skapas enligt metod 1.

2.4.1 Förväntat antal kanter och trianglar

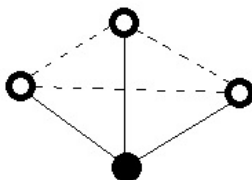
Om vi antar att vi har en graf $G(n, p)$ där $p = \frac{\lambda_1}{n-1}$ har vi att det förväntade antalet kanter är $\frac{n\lambda_1}{2}$, och det förväntade antalet trianglar är approximativt $\frac{\lambda_1^3}{6}$. För att öka grafens transitivitet tar vi för varje öppen tripplett centrerad i någon nod v och lägger till den triangelskapande kanten med en sannolikhet q_1 . Hur många öppna trippletter har vi då i vår graf $G(n, p)$? Till att börja med konstaterar vi att en öppen tripplett centrerad i någon nod v medför att nodens grad d är större än 1. Antalet öppna trippletter centrerade i nod v är antalet olika par vi kan välja ut bland grannarna, förutsatt att vi inte har några kanter mellan nod v 's grannar. Risken att detta skulle inträffa går mot 0 då $n \rightarrow \infty$ och vi har därför att antalet öppna trippletter centrerade i nod v är $\binom{d}{2} = \frac{d(d-1)}{2}$. Enligt definitionen av väntevärde får vi då att $E[\text{antal öppna trippletter centrerade i nod } v] = \sum_{i=2}^{n-1} P(i \text{ grannar}) \frac{i(i-1)}{2}$. Antalet grannar vet vi sedan tidigare är binomi-

alfördelat och kan approximeras med en poissonfördelning med parameter λ_1 . Om vi låter T_v vara antalet öppna tripletter centrerade i nod v ger detta att

$$E[T_v] = \sum_{i=0}^{n-1} \frac{\lambda_1^i}{i!} e^{-\lambda_1} \frac{i(i-1)}{2} = \frac{\lambda_1^2}{2} \sum_{j=0}^{n-1} \frac{\lambda_1^j}{j!} e^{-\lambda_1} = \frac{\lambda_1^2}{2}.$$

Vi kan alltså förvänta oss $\frac{n\lambda_1^2}{2}$ öppna tripletter i vår graf och om var och en av dessa tripletter oberoende bildar en triangel med en sannolikhet q_1 får vi att förväntat antal tillagda trianglar är $\frac{q_1 n \lambda_1^2}{2}$. Detta förutsätter dock att inga öppna tripletter har identiska slutnoder då detta skulle medföra att den potentiellt triangelskapande kanten är densamma för båda tripletterna. Men då n blir stort samtidigt som det förväntade gradtalet är ändligt går denna sannolikhet mot 0. Det finns dock ett problem. Om vi tänker oss att en nod har 3 grannar och alla 3 triangelskapande kanter adderas, skapar vi faktiskt fyra trianglar istället för 3, se figur 2. vilket sker med en sannolikhet q_1^3 . Vi kommer att bortse från detta fortsättningsvis då problemet är relativt litet. Vår slutliga graf har således ett förväntat antal trianglar är approximativt $\frac{3q_1 n \lambda_1^2 + \lambda_1^3}{6}$.

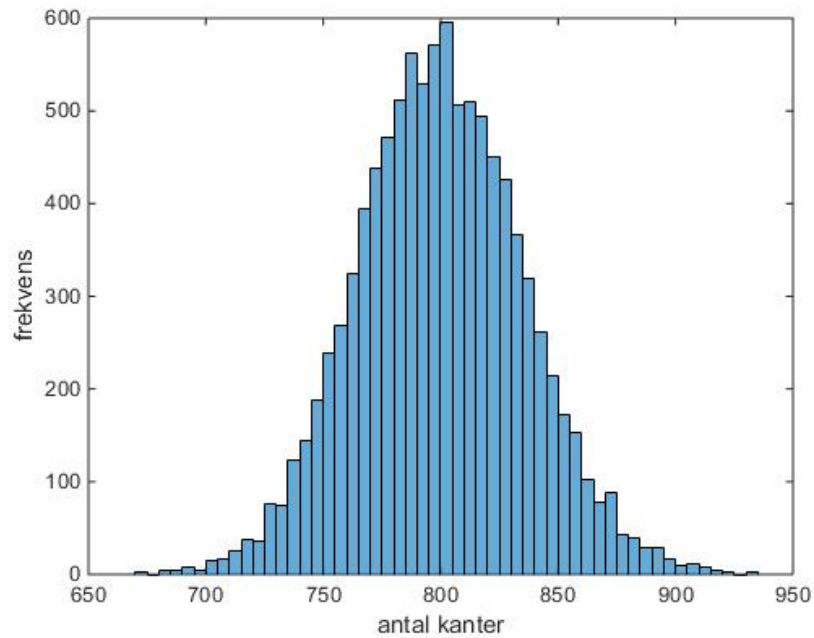
Figur 2: Om de 3 öppna tripletterna alla skulle stängas med varsin kant (de streckade linjerna), skapas 4 trianglar och inte 3.



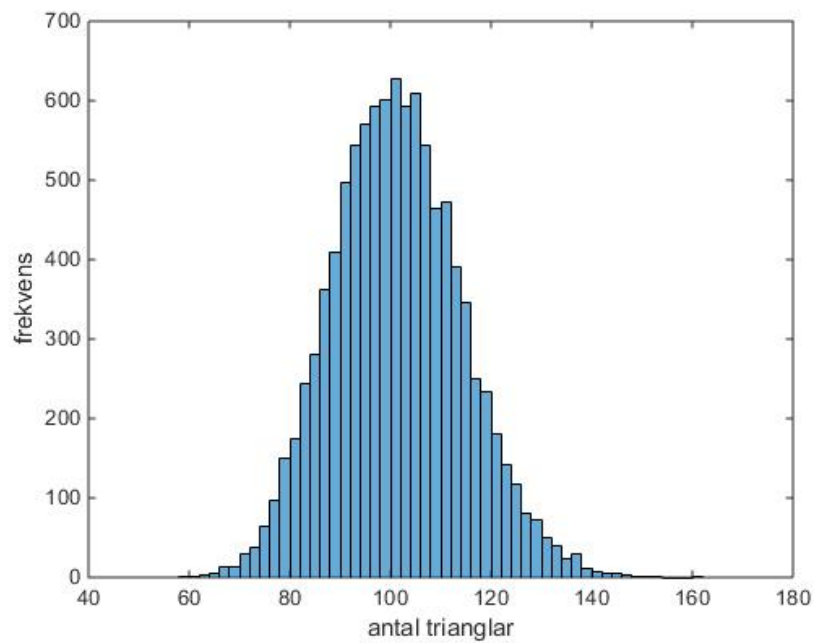
Vad gäller det förväntade antalet kanter har vi till att börja med alltså $\frac{n\lambda_1}{2}$ kanter från $G(n, p)$. Då vi förväntar oss $\frac{n\lambda_1^2}{2}$ öppna tripletter kan vi förvänta oss att vi lägger till $\frac{q_1 n \lambda_1^2}{2}$ kanter. Det förväntade totala antalet kanter blir då genom addition $\frac{n(\lambda_1 + q_1 \lambda_1^2)}{2}$.

Låt oss nu titta på en simulering gjord i matlab. Om vi låter skapa 10000 grafer med 1000 noder enligt metod 1, samtliga med parametrarna $\lambda_1 = 1.3986$ och $q_1 = 0.1023$ har vi enligt formlerna ovan att det förväntade antalet kanter och trianglar är 800 respektive 101. Räknar vi det totala antalet kanter och trianglar kan vi se fördelningarna i figur 3 och 4.

Figur 3: Histogram över fördelningen av antalet kanter. 10000 simuleringar av en graf med 1000 noder och parametrar $\lambda_1 = 1.3986$ och $q_1 = 0.1023$.



Figur 4: Histogram över fördelningen av antalet trianglar. 10000 simuleringar av en graf med 1000 noder och parametrar $\lambda_1 = 1.3986$ och $q_1 = 0.1023$.



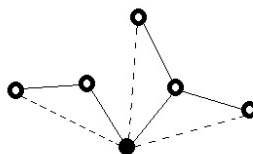
Då vi ser att histogrammen verkar centrerade runt värdena 800 och 100 stämmer simuleringarna väl överens med formlerna. Att fördelningen liknar en normalfördelning följer av att en poissonfördelning kan approximeras med en normalfördelning vid små λ .

2.4.2 Gradfördelning

Gradfördelningen i en graf skapad enligt metod 1 går eventuellt att bestämma analytiskt men faller utanför ramen inom detta arbete. Den förväntade graden däremot ska vi ta reda på nu.

Om vi låter D_v vara graden för nod v innan vi lagt till några extra kanter så vet vi sedan tidigare att $D_v \sim Po(\lambda_1)$ och dess realisering skriver vi d_v . Vidare låter vi $D_{v_1}, D_{v_2}, \dots, D_{v_{d_v}}$ vara grannarnas utökade grader (eng. excess degree) där vi bortser från kanten till nod v . Dessa är då alla asymptotiskt oberoende och $Bin(n-2, \frac{\lambda_1}{n-1})$ -fördelade. Denna fördelning kan approximeras som $Po(\lambda_1)$. Genom att sätta en kant mellan nod v och dess grannars grannar kan vi nu skapa en triangel, varför de adderas med en sannolikhet q_1 och på så sätt ökar graden för nod v , se figur 5.

Figur 5: De streckade linjerna i figuren är potentiella triangelskapande kanter vilka om de adderas ökar graden för nod v .



Vi låter nu X = antalet adderade kanter och får att

$$E[X|D_v = d_v] = E[(D_1 + D_2 + \dots + D_{d_v})q_1] = \lambda_1 q_1 d_v.$$

Detta ger vidare att

$$E[X] = E[E[X|D_v]] = E[\lambda_1 q_1 D_v] = \lambda_1^2 q_1.$$

Den förväntade totala graden $E[D_{tot}]$ efter att potentiella triangelskapande kanter adderats blir genom lineariteten hos väntevärdet slutligen $\lambda_1^2 q_1 + \lambda_1$.

Centrala slutsatser då trianglar adderas enligt metod 1:

- $E[\text{antal kanter}] = \frac{n(\lambda_1 + q_1 \lambda_1^2)}{2}$
- $E[\text{antal trianglar}] = \frac{3q_1 n \lambda_1^2 + \lambda_1^3}{6}$
- $E[D_{tot}] = \lambda_1^2 q_1 + \lambda_1$

2.5 Triangelutökning enligt metod 2

Att utöka en ER-graf med ytterligare trianglar enligt metod 2 går ut på att först skapa en triangelgraf bestående av slumpmässigt utplacerade trianglar för att sedan skapa unionen av triangelgrafen och ER-grafen, där kanter som förekommer flera gånger reduceras till en. Vi börjar återigen med att undersöka vad det förväntade antalet kanter och trianglar är.

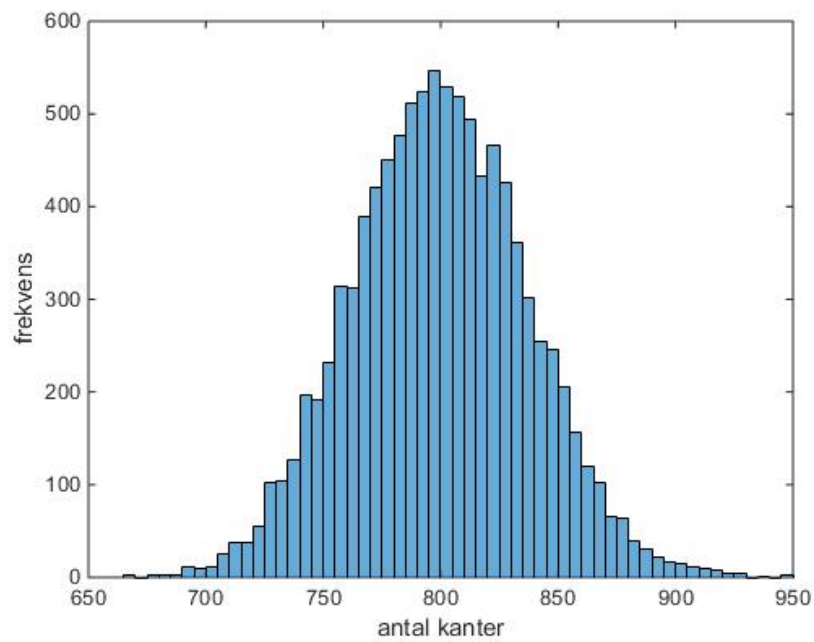
2.5.1 Förväntat antal kanter och trianglar

Om vi antar att vi har en graf $G(n, p)$ där $p = \frac{\lambda_2}{n-1}$ har vi att det förväntade antalet kanter är $\frac{n\lambda_2}{2}$, och det förväntade antalet trianglar är approximativt $\frac{\lambda_2^3}{6}$. För att öka grafens transitivitet lägger vi till, oberoende av varandra, var och en av grafens $\binom{n}{3}$ möjliga trianglar med en sannolikhet r . Förväntat antal tillagda trianglar blir därmed $\binom{n}{3}r = \frac{n(n-1)(n-2)r}{6}$. Notera att detta endast gäller då $n \rightarrow \infty$ ty annars kommer vi att få fler trianglar än bara de vi placerar ut då trianglar skulle kunna skapas "utav misstag". För att antalet adderade trianglar skall vara av samma storleksordning som i den tidigare metoden sätter vi $r = \frac{q_2}{n^2}$ vilket ger att det förväntade antalet tillagda trianglar är approximativt $\frac{nq_2}{6}$. Det totala antalet trianglar i den nya grafen blir således approximativt $\frac{\lambda_2^3 + nq_2}{6}$ om vi tänker att sannolikheten att vi placerar en ny triangel på en redan existerande är nära 0 för stora n .

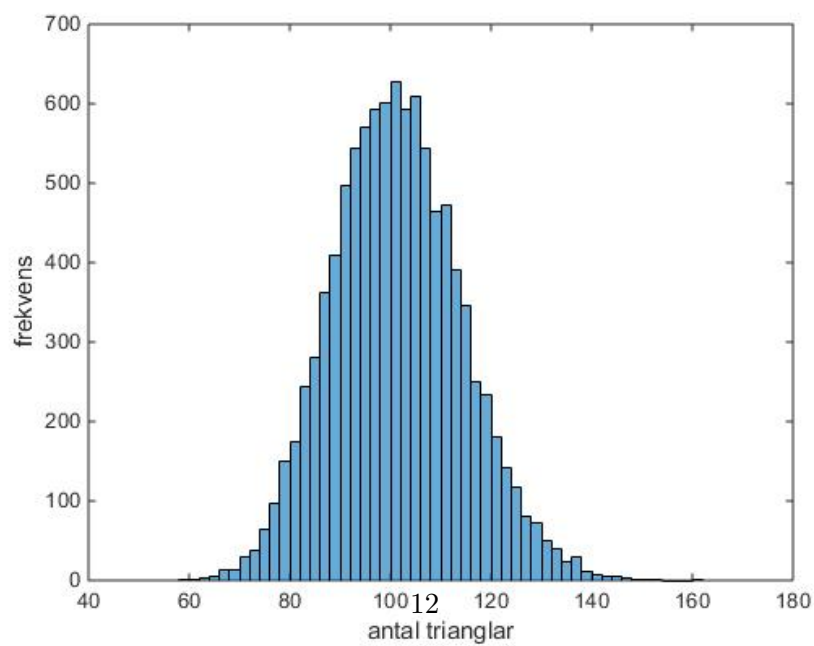
Vad gäller det totala antalet kanter har vi att det förväntade antalet kanter från den ursprungliga grafen $G(n, p)$ är $\frac{n\lambda_2}{2}$. Om vi nu adderar $\frac{nq_2}{6}$ nya trianglar innebär det att vi adderar $3\frac{nq_2}{6}$ nya kanter. Frågan är nu hur många dubbelkanter vi skapar när vi placerar trianglarna. Vi kan konstatera att andelen redan upptagna kanter är $\frac{n\lambda_2}{2} / \binom{n}{2} = \frac{\lambda_2}{n-1}$ vilket är nära 0 då n är stort. Vi kan därför anta att vi inte råkar ut för några dubbelkanter asymptotiskt och konstatera att det förväntade antalet kanter blir $\frac{n\lambda_2}{2} + 3\frac{nq_2}{6} = \frac{n(\lambda_2 + q_2)}{2}$.

Låt oss nu titta på en simulering gjord i matlab. Om vi låter skapa 10000 grafer med 1000 noder enligt metod 2, samtliga med parametervärdena $\lambda_2 = 0.9990$ och $q_2 = 0.6018$ har vi enligt formlerna ovan att det förväntade antalet kanter och trianglar är 800 respektive 101 precis som i avsnitt 2.4.1. Räknar vi det totala antalet kanter och trianglar kan vi se fördelningarna i figur 6 respektive 7.

Figur 6: Histogram över fördelningen av antalet kanter. 10000 simuleringar av en graf med 1000 noder och parametervärden $\lambda_2 = 0.9990$ och $q_2 = 0.6018$.



Figur 7: Histogram över fördelningen av antalet trianglar. 10000 simuleringar av en graf med 1000 noder och parametervärden $\lambda_2 = 0.9990$ och $q_2 = 0.6018$.



Då vi ser att histogrammen verkar centrerade runt värdena 800 och 100 stämmer simuleringarna väl överens med formlerna. Att fördelningen liknar en normalfördelning följer som tidigare av att en poissonfördelning kan approximeras med en normalfördelning vid små λ .

2.5.2 Gradfördelning

Tidigare har vi konstaterat att gradfördelningen för en given nod v i en ER-graf är $Po(\lambda_2)$. Låt oss skapa en ny graf med samma dimension som ER-grafen där vi inkluderar var och en av de $\binom{n}{3}$ trianglarna med någon sannolikhet y som tidigare fastställts till $\frac{q_2}{n^2}$. Då har vi att graden för en given nod $v = 2X$, där X är antalet trianglar med ett hörn i v . Då det finns $\binom{n-1}{2}$ olika trianglar med ett hörn i v måste X vara $Bin(\binom{n-1}{2}, \frac{q_2}{n^2}) \approx Po(q_2/2)$. Detta ger att en given nod v kommer att vara $2Po(q_2/2)$ -fördelad, se [3]. Fördelningen har samma väntevärde som en $Po(q_2)$ -fördelning men dubbelt så stor varians och egenskapen att den antar udda heltal med sannolikheten 0.

Om vi nu slår ihop graferna och antar att inga dubbelkanter inträffar så blir den totala graden, D_{tot} , för nod v $(2Po(q_2/2) + Po(\lambda_2))$ -fördelad med väntevärde $q_2 + \lambda_2$

Centrala slutsatser då trianglar adderas enligt metod 2:

- $E[\text{antal kanter}] = \frac{n(\lambda_2 + q_2)}{2}$
- $E[\text{antal trianglar}] = \frac{\lambda_2^3 + nq_2}{6}$
- $D_{tot} \sim 2Po(q_2/2) + Po(\lambda_2)$

3 En jämförelse av den största komponenten hos de två metoderna

För att två noder ska tillhöra samma komponent krävs att det existerar en väg av kanter mellan dem. En känd egenskap hos ER-grafen, visad av Erdős år 1959, är att den genomgår en så kallad fasövergång. I [5, kap. 4] visas att

1. Då $\lambda < 1$ kommer grafen ha många små komponenter, den största av ordning $\log(n)$.
2. Då $\lambda > 1$ kommer grafen med sannolikhet 1 ha en jättekompont innehållandes en positiv andel av noderna.

Vi kommer här inte att teoretiskt försöka bestämma storleken på den största komponenten efter att trianglar adderats. Istället nöjer vi oss med att med hjälp av simulering jämföra de två metoderna. Med största komponenten, K , menar vi fortsättningsvis andelen noder tillhörandes den största

komponenten. För att kunna göra en meningsfull jämförelse kommer vi att hålla det förväntade antalet kanter och trianglar jämbördigt mellan de två metoderna. Då vi har slutna uttryck för dessa storheter innehållandes parametrarna $\lambda_1, q_1, \lambda_2, q_2$ kan detta ordnas genom kalibrering av parametrarna baserat på ekvationssystemet

$$\begin{cases} \frac{n(\lambda_2+q_2)}{2} = \frac{n(\lambda_1+q_1\lambda_1^2)}{2} & (\text{förväntat antal kanter}) \\ \frac{\lambda_2^3+nq_2}{6} = \frac{3nq_1\lambda_1^2+\lambda_1^3}{6} & (\text{förväntat antal trianglar}) \end{cases}$$

vilket då vi låter n gå mot oändligheten kan förenklas till

$$\begin{cases} \lambda_2 + q_2 = \lambda_1 + q_1\lambda_1^2 \\ q_2 = 3q_1\lambda_1^2. \end{cases}$$

Vi börjar med att konstatera att om förväntade antalet trianglar hålls litet kommer båda metoderna att närma sig en helt vanlig ER-graf och vi har inget intressant att studera. Vidare inses att med metod 1 kommer adderade triangelskapande kanter inte att påverka den största komponenten i grafen. Att lägga till en kant mellan två noder som redan är sammankopplade på omvägar ändrar inget, och storleken på dess största komponent kommer enbart att bero på parametern λ_1 identiskt med en enkel ER-graf. Att slumpmässigt lägga ut trianglar bör däremot ha en påverkan. En utförlig jämförelse skulle innebära att jämföra två ytor med varandra. Även om detta är möjligt nöjer vi oss här med en mer anekdotisk jämförelse vid ett antal punkter, se tabell 1.

Vi ser att de största komponenterna ligger en bit ifrån varandra för vissa kombinationer av förväntat antal kanter och trianglar och närmre för andra. För att få en uppfattning om den förväntade största komponenten skiljer sig för metoderna skapar vi nu två konfidensintervall, ett för metod 1 och ett för metod 2, och ser om de överlappar varandra. Vi väljer att göra detta på nivån där vi håller det förväntade antalet kanter på 8500 och det förväntade antalet trianglar på 1667. Då vi inte har någon fördelning får vi ett approximativt intervall för väntevärdena enligt centrala gränsvärdessatsen om vi gör 35 simuleringar och skattar variansen med stickprovsvariansen enligt formeln

$$1 - \alpha \approx P\left(\bar{K} - \frac{Sz_{\alpha/2}}{\sqrt{n}} < \mu < \bar{K} + \frac{Sz_{\alpha/2}}{\sqrt{n}}\right).$$

Med en konfidensgrad på 0.99 blir resultatet för metod 1 (0.4780, 0.4954) och för metod 2 (0.5310, 0.5422). Dessa intervall överlappar inte varandra så vi har ett resultat som tyder på att största komponentens storlek skiljer sig åt för metoderna och därmed att fasövergångarna inte är identiska. Vidare kan vi i tabell ?? se att då vi håller antalet kanter och trianglar lågt tenderar metod 1 att generera en liten största komponent jämfört med metod 2, och tvärtom för ett högre antal kanter och trianglar. Hur detta kan förklaras

Tabell 1: Jämförelse av den största komponenten (K) där $\bar{K}_1$ är medelvärdet av 5 simuleringar med metod 1 och $\bar{K}_2$ är medelvärdet av 5 simuleringar med metod 1. Antalet noder i de simulerade graferna är 10000.

E[antal kanter]	E[antal trianglar]	$\bar{K}_1$	$\bar{K}_2$
5041	347	0.0231	0.0530
5563	355	0.0658	0.1389
5800	333	0.1683	0.2162
6101	451	0.2288	0.2755
6443	315	0.3266	0.3666
7000	167	0.4796	0.4849
8500	1667	0.4810	0.5409
9600	1667	0.6336	0.6393
8749	417	0.6786	0.6710
9999	834	0.7473	0.7377
12 642	2381	0.8107	0.7885
21 266	5556	0.9489	0.9148

är inte genast uppenbart och skulle kunna vara en fråga för vidare studier. Det är tänkbart att då vi håller antalet trianglar högt i förhållande till antalet kanter kommer metoderna att glida ifrån varandra med avseende på storleken av deras största komponent. Om vi t.ex. gör en simulering där det förväntade antalet kanter och trianglar till 7443 respektive 2381 får vi att medelvärdet av den största komponenten med metod 1 är 0.0542 och hela 0.3498 med metod 2.

4 En jämförelse av klusterkoefficienten hos de två metoderna

Klustringen i en slumpgraf är ofta en intressant egenskap att undersöka. Vi ska i följande kapitel se att den förväntade klusterkoefficienten går att bestämma då vi utökar en ER-graf med trianglar enligt metod 2, men att vi stöter på problem om vi utökar den med trianglar enligt metod 1.

4.1 Förväntad klusterkoefficient med metod 2

Klustercoefficienten C definierades i avsnitt 2.3 men en ekvivalent och nu mer användbar definition är

$$C = \frac{\frac{1}{n} \sum_{i=1}^n \text{antalet ordnade stängda tripletter centrerade hos nod } v_i}{\frac{1}{n} \sum_{i=1}^n \text{antalet ordnade tripletter centrerade hos nod } v_i}.$$

Vi betecknar storheterna i täljaren och nämnaren med A_n respektive B_n . Vi har tidigare visat att antalet trianglar med ett hörn i en given nod v är $Po(q_2/2)$ -fördelat. Detta medför att antalet ordnade stängda tripletter i en given nod v är $2 * Po(q_2/2)$ -fördelat och alltså har väntevärdet q . Stora talens lag ger därför att $A_n \rightarrow q_2$ i sannolikhet då antalet ordnade stängda tripletter centrerade i de olika noderna är asymptotiskt oberoende.

Antalet sammanbundna tripletter centrerade i en given nod är $\binom{d}{2}$ där d är nodens grad. Antalet ordnade sammanbundna tripletter centrerade i en given nod är därför $2\binom{d}{2} = d(d-1)$. D vet vi är $(2Po(\frac{q_2}{2}) + Po(\lambda_2))$ -fördelat och då blir $E[D(D-1)] = E[D^2 - D] = Var[D] + E[D]^2 - E[D]$ vilket med algebraiska standardmetoder blir $q_2 + (q_2 + \lambda_2)^2$. Stora talens lag ger därför att $B_n \rightarrow q_2 + (q_2 + \lambda_2)^2$ i sannolikhet då antalet ordnade tripletter centrerade i de olika noderna är asymptotiskt oberoende.

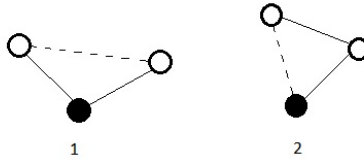
Slutskys sats, se [4, s. 168], ger oss nu att $\frac{A_n}{B_n} \rightarrow \frac{q_2}{q_2 + (q_2 + \lambda_2)^2}$ i fördelning, och då det är en degenererad variabel ger sats 3.2 i [4, s. 157] att konvergensten också sker i sannolikhet. Den förväntade klusterkoefficienten är alltså $\frac{q_2}{q_2 + (q_2 + \lambda_2)^2}$.

4.2 Förväntad klusterkoefficient med metod 1

Om vi nu utgår från en ER-graf och bortser från de få trianglar som skapats kan vi konstatera att en given nod v kan komma att utgöra ett hörn i en triangel på två sätt.

1. En kant dras mellan nodens grannar.
2. En kant dras mellan noden och en grannes granne, se figur 8.

Figur 8: Två sätt en nod kan komma att utgöra ett hörn i en triangel



För att bestämma täljaren A_n i klusterkoefficienten behöver vi hitta det förväntade antalet stängda tripletter centrerade i nod v . Om vi börjar med fall 1, kanter som dras mellan nodens grannar så har vi tidigare visat att $D \sim Po(\lambda_1)$ och det finns $\binom{d}{2}$ möjliga triangelskapande kanter. Var och en av dessa adderade kanter skapar nu 2 ordnade stängda tripletter centrerade

i v , dvs $d(d-1)$ stycken. Vi har att

$$E[q_1 D(D-1)] = q_1 \sum_{k=0}^{\infty} k(k-1) \frac{e^{-\lambda_1} \lambda_1^k}{k!} = q_1 \lambda_1^2 \sum_{l=0}^{\infty} \frac{e^{-\lambda_1} \lambda_1^l}{l!} = \lambda_1^2 q_1.$$

För fall 2 visade vi i avsnitt 2.4.2 att det förväntade antalet kanter adderade mellan en nod och dess grannars grannar är $\lambda_1^2 q_1$. Då var och en av dessa bidrar med 2 ordnade stängda tripletter får vi att det totala förväntade antalet ordnade stängda tripletter i nod v är $3\lambda_1^2 q_1$. Enligt stora talens lag går därför $A_n \rightarrow 3\lambda_1^2 q_1$ i sannolikhet.

Vad gäller nämnaren B_n i klusterkoefficienten kan vi börja med att konstatera att vi kan förvänta oss λ_1^2 ordnade öppna tripletter centrerade i v från ER-grafen, enligt avsnitt 2.4.1. Vidare noterar vi att adderandet av en kant mellan 2 grannar till nod v inte förändrar antalet ordnade tripletter centrerade i nod v (en öppen byts mot en stängd). En kant mellan v och en grannes granne däremot ökar antalet ordnade tripletter centrerade i v . Efter att samtliga dessa kanter har adderats är nodens totala grad $D_{tot} = N + D$ där N är antalet adderade kanter. Vi har därför skapat $\binom{D_{tot}}{2} - N$ nya öppna tripletter vilket innebär $D_{tot}(D_{tot} - 1) - 2N$ nya ordnade öppna tripletter. Väntevärdet av denna stokastiska variabel är svår att beräkna då det inbegriper beräkning av $E[D_{tot}^2]$. Det är tänkbart att andra mått på klustringen i grafen skulle fungera bättre. $P[v_i, v_j \text{ är grannar} | v_i, v_j \text{ är båda grannar till } v_k]$ är ett sådant tänkbart mått. En sådan undersökning faller dock utanför omfånget för denna text.

5 Diskussion

I denna uppsats har vi tittat på två olika metoder att öka transitiviteten i en ER-graf $G(n, p)$. Båda metoderna är slumpmässiga till sin natur och metod 1 bygger på att en kant adderas till varje öppen tripplett med någon sannolikhet för att på så sätt skapa trianglar. Metod 2 bygger på att en ny triangelgraf skapas genom att placera ut var och en av de möjliga trianglarna med någon sannolikhet och sedan skapas unionen av denna triangelgraf och ER-grafen där eventuella dubbelkanter tas bort.

I vår analys av grafernas har vi utgått från en graf där antalet noder är mycket stort och med den utgångspunkten började vi att med analytiska resonemang finna slutna formler som med hjälp av två parametrar bestämmer det förväntade antalet kanter och trianglar. För metod 2 har vi vidare lyckats bestämma grafens gradfördelning och för metod 1 har vi bestämt den förväntade graden.

En annan intressant egenskap vi studerat är storleken på grafernas största komponent. Då detta är svårt att analysera teoretiskt har vi med hjälp av grafsimuleringar kunnat visa att även om metoderna har samma förväntade antal kanter och trianglar är metoderna inte ekvivalenta vad gäller storleken

på den största komponenten. Detta skulle i så fall betyda att fasövergångarna inte är identiska och följdaktligen att metoderna är fundamentalt olika. Vi har kunnat visa att då det förväntade antalet kanter och trianglar är litet tenderar metod 1 att ge en mindre största komponent jämfört med metod 2, och att det omvända gäller då vi har ett större antal kanter och trianglar. Vi har också påvisat att det finns extremfall där metoderna ger väldigt olika resultat och en hypotes är att extremfallen har någon koppling till proportionerna mellan det förväntade antalet kanter och trianglar i en graf. Det skulle således vara intressant att undersöka och jämföra metodernas fasövergångar ytterligare och kanske speciellt var vi kan förvänta oss hitta dessa extremfall.

Referenser

- [1] Bollobas, B. (2001): Random graphs. Cambridge University Press.
- [2] Britton, T., och Alm, S. E. (2011): Stokastik - Sannolikhetssteori och statistikteori med tillämpningar, Liber.
- [3] Deijfen, M. (2004): Epidemispredning på sociala grafer, Stockholms Universitet.
- [4] Gut, A. (2009): An intermediate course in probability, Springer.
- [5] Hofstad, R. (2014): Random graphs and complex networks, Eindhoven university of technology.
- [6] Newman M. E. J., Strogatz S. H. and Watts D. J.: Random graphs with arbitrary degree distributions and their applications, Physical Review E, SIAM 64:026118