Il fenomeno Small World

Capitolo 20 del testo

Esperimento di Milgram

- Lo psicologo sociale Stanley Milgram nel 1967 condusse il seguente esperimento:
  - scelse una persona – un *destinatario*, una persona alla quale doveva essere recapitata una lettera della quale lo stesso Milgram era il mittente
  - per fissare le idee, l'obiettivo risiedeva molto lontano da Milgram – ad esempio, Milgram e l'obiettivo erano sulle coste opposte degli Stati Uniti.
  - Milgram ha poi scelto a caso un insieme di iniziatori e ha consegnato ad ognuno di essi una copia della lettera
  - ha anche fornito ad ogni iniziatore una serie di informazioni sul destinatario: nome, indirizzo, occupazione, e informazioni personali quali interessi, passatempi, …
  - Milgram ha chiesto ad ogni iniziatore di fare in modo di far giungere la copia della lettera in suo possesso al destinatario, senza, però, inviargliela direttamente a mezzo sistema postale
  - invece, ogni iniziatore doveva: scrivere il suo nome sulla lettera e, *con l'obiettivo di far giungere la lettera al destinatario nel minor numero di passi possibile*, consegnarla (o inviarla) a un suo diretto conoscente chiedendogli di ripetere le medesime azioni

Esperimento di Milgram

- E cosa ha osservato Stanley Milgram al termine del suo esperimento?
  - innanzi tutto, che circa un terzo delle lettere hanno raggiunto il destinatario
  - poi, che le lettere che hanno raggiunto il destinatario lo hanno fatto, in media, in sei passi
- Le lettere hanno viaggiato da una parte all'altra degli Stati Uniti, trasmesse da un individuo all'altro sulla base di conoscenze personali, e hanno impiegato in media 6 passi per giungere a destinazione!
  - I "famosi" 6 *gradi di separazione*

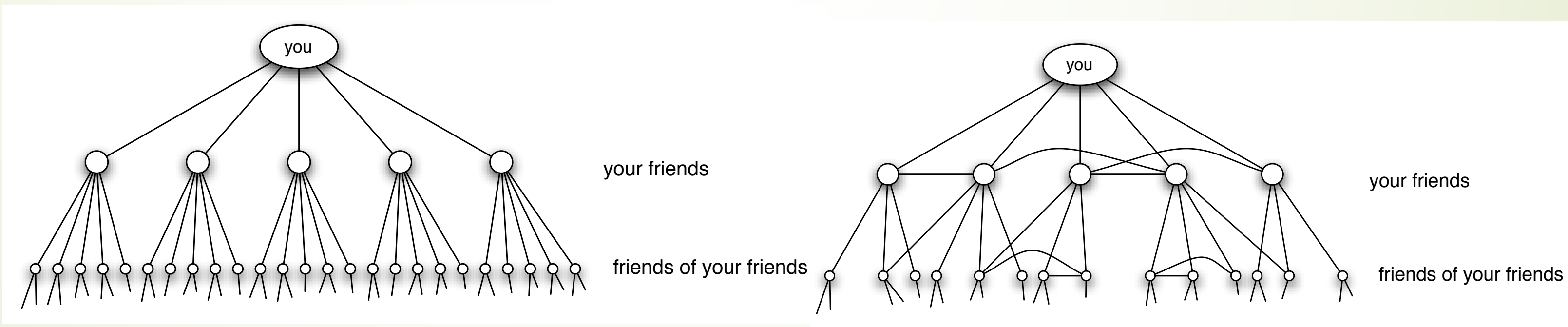
- Ebbene, in base all'esito dell'esperimento di Milgram, possiamo trarre due conclusioni:
  - 1) in una rete sociale è presente una moltitudine di percorsi molto brevi, che connettono qualunque coppia di nodi – il fenomeno **Small World**
  - 2) che i percorsi brevi non solo esistono, ma *possono essere trovati con facilità*
    - ossia, da nodi che non conoscono altro della struttura della rete se non i propri immediati vicini!

1) Una moltitudine di percorsi brevi

- Ci domandiamo ora: quale spiegazione intuitiva possiamo trovare al fatto che in una rete sociale esistano tanti shortest paths fra una coppia di nodi?
- Intuitivamente, possiamo vederla così:
  - io ho, diciamo, 100 amici
  - ciascuno dei quali ha, diciamo, 100 amici
  - ciascuno dei quali ha 100 amici, e così via
  - questo significa che il grafo delle relazioni in questo gruppo di persone contiene 10000 percorsi di lunghezza 2 da me ad altre persone della rete – 1000000 percorsi di lunghezza 3 da me ad altre persone della rete
  - cioè, io sono collegata a mezzo di percorsi molto brevi a un sacco di gente!
  - E, poiché posso ripetere lo stesso ragionamento per qualunque altro individuo che popola la rete, ecco la spiegazione dell'esistenza di tanti percorsi brevi in una rete sociale
- Bello, intuitivo, facile, ma questo ragionamento ha una pecca…

Una moltitudine di percorsi brevi

- Bello, intuitivo, facile, ma il ragionamento che abbiamo illustrato ha una pecca: non tiene conto della ***chiusura triadica*** (della quale avremo modo di parlare in seguito)
  - ossia, del fatto che in una rete sociale esistono tanti triangoli
  - ossia, se un individuo a conosce due individui b e c, allora è probabile che prima o poi anche b e c si conosceranno
  - perché, ad esempio, i miei amici, frequentandomi, avranno possibilità di incontrarsi!
- Quindi, fra i 100 amici dei miei amici, si troveranno anche alcuni dei miei amici
  - e il grafo della relazione assomiglierà alla figura a destra, piuttosto che a quella a sinistra



(a) *Pure exponential growth produces a small world*

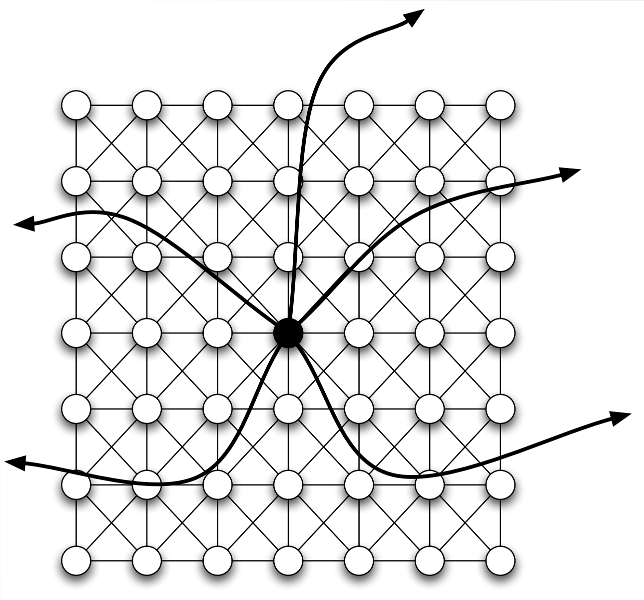
(b) *Triadic closure reduces the growth rate*

Il modello Watts-Strogatz

- Ciò premesso, ci proponiamo di studiare un modello generativo di grafi aleatori che generi
  - a) Small Worlds
  - b) contenenti molte chiusure triadiche
- Il modello proposto da **Watts** e **Strogatz** (1998) consiste di un grafo fissato deterministicamente ed un insieme di archi casuali
- Il grafo fissato deterministicamente è una griglia "arricchita" – che corrisponde, sostanzialmente a uno Unit Disk Graph
  - considerando i nodi come punti di uno spazio metrico bidimensionale
  - i nodi sono disposti sui punti a coordinate intere di un quadrato centrato nell'origine degli assi cartesiani
  - e ogni nodo è collegato a ciascuno dei nodi vicini in orizzontale, verticale e diagonale
  - Formalmente: fissato $n \in \mathbb{N}$, $V = \{ (i, j): 0 \leq i \leq n \wedge 0 \leq j \leq n \}$ e ciascun nodo (i,j) con $0 < i < n$ e $0 < j < n$ è collegato ai nodi $(i,j-1)$, $(i+1, j-1)$, $(i+1,j)$, $(i+1,j+1)$, $(i,j+1)$, $(i-1, j+1)$, $(i-1, j)$ $(i-1, j-1)$ e analogamente per i nodi (i,j) con $i \in \{0, n\}$ e/o $j \in \{0, n\}$

Il modello Watts-Strogatz

- Il modello proposto da **Watts** e **Strogatz** (1998) consiste di un grafo fissato deterministicamente ed un insieme di archi casuali
- Il grafo fissato deterministicamente è una griglia "arricchita"
  - ossia, i nodi sono disposti sui punti a coordinate intere di un quadrato centrato nell'origine degli assi cartesiani e ogni nodo è collegato a ciascuno dei nodi vicini in orizzontale, verticale e diagonale



- Poi, fissato un valore k, ogni nodo sceglie uniformemente a caso k nodi che diventeranno suoi vicini
- In effetti, comunque, più che a una griglia su una superficie piana, dobbiamo pensare a *una griglia "appoggiata" su una superficie sferica*
  - che si "richiude" su sé stessa
  - che chiameremo *wrapped*

Il modello Watts-Strogatz

- Osserviamo, ora, un grafo generato in accordo al modello di Watts-Strogatz
- 1) possiamo individuare una sorta di dicotomia relazioni locali / relazioni a distanza soggiacente fra gli archi deterministici e quelli random
  - gli archi della griglia, che costituiscono "l'ossatura fissa" del grafo, rappresentano le relazioni fra nodi "fisicamente" vicini - quelli le cui coordinate differiscono di poco
  - gli archi random esprimono relazioni fra nodi "fisicamente" lontani
- 2) possiamo ben immaginare che i nodi "fisicamente" vicini abbiano più probabilità di incontrarsi rispetto ai nodi "fisicamente" lontani
  - ossia, che i nodi vicini abbiano frequentazioni assidue, quelli lontani no
  - possiamo pensare, allora, agli archi della griglia come archi che rappresentano relazioni forti (*strong ties*)
  - e agli archi random come archi che rappresentano relazioni deboli (*weak ties*)
- 3) i triangoli sono sempre presenti a livello locale e sono poco probabili fra gli archi random
  - confermando l'idea di cui al punto 2): infatti, i triangoli si formano quando individui che hanno un amico comune si incontrano, e due amici di uno stesso individuo che vivono ai due capi opposti della Terra non è probabile che abbiano molte occasioni per incontrarsi e diventare amici

Il modello Watts-Strogatz

- Dunque, un grafo generato dal modello Watts-Strogatz contiene molti triangoli
  - in accordo a quel che si riscontra, generalmente, nelle reti sociali
- Rimane l'altra questione: sarà vero che coppie di nodi qualunque sono collegati da numerosi percorsi brevi?
- A tal proposito, Watts e Strogatz hanno osservato che:
  - se partiamo da un nodo u e
  - a partire da u, per un certo numero di passi ci muoviamo lungo gli archi random,
  - poiché gli archi random sono distribuiti uniformemente nel grafo, è molto improbabile che, in questo procedimento tocchiamo due volte lo stesso nodo
  - ossia, molto probabilmente, in h passi abbiamo la possibilità di raggiungere k^h nodi
- Il ragionamento di Watts-Strogatz, appena descritto, è basato su considerazioni intuitive
- Successivamente, Bollobás e Chung (1988) hanno formalmente dimostrato questo punto
  - e hanno anche individuato la lunghezza media degli shortest paths nei grafi generati in accordo al modello di Watts-Strogatz

Il modello Watts-Strogatz

- Nel modello di Watts-Strogatz da ogni nodo partono k archi random

- In effetti, comunque, il ragionamento intuitivo di Watts e Strogatz può essere ripetuto su un modello in cui è presente di gran lunga meno casualità:
- è sufficiente che **soltanto da un nodo su k partano archi random** e che, inoltre **da tale nodo parta un solo arco random**
  - Idea della motivazione: raggruppiamo quadrati di $k \times k$ nodi della griglia in città
    - dove ogni città ha uno e un solo arco random uscente
  - e ripetiamo il ragionamento sopra a livello di città: in h passi possiamo giungere in k^h città,
  - infine all'interno di una città ci muoviamo attraverso gli archi della griglia

- Possiamo, quindi, concludere che

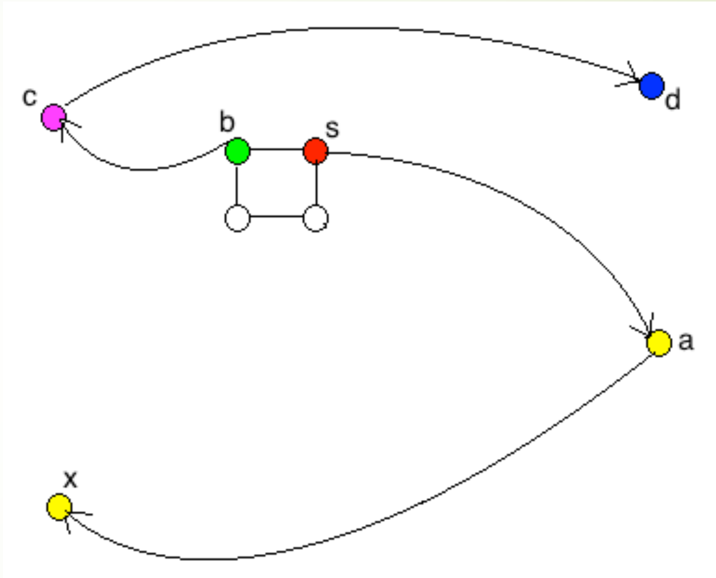
poca casualità è sufficiente per avere tanti shortest paths

2) Percorsi brevi facili da trovare

- Ora: siamo il nodo u in un grafo di Watts-Strogatz e vogliamo inviare un messaggio ad una certa destinazione v **di cui conosciamo le coordinate**
  - e, dalle sue coordinate, sappiamo che v è molto distante da noi
- naturalmente, cerchiamo di fare arrivare il messaggio il più velocemente possibile
- ossia, cerchiamo di fare in modo che venga consegnato attraverso uno shortest path (che, probabilmente, sarà breve) che collega u (noi) a v
- ma noi (che siamo il nodo u), meschini, non conosciamo altro, della rete, che i nostri contatti
  - ossia, i nodi ai quali siamo collegati – i nostri vicini nel grafo
- allora, facciamo tante copie del messaggio quanti sono i nostri vicini e mandiamo una copia a ciascun vicino
  - tanto, abbiamo soltanto 8/9 vicini – non andiamo falliti in fotocopie
- non possiamo fare altro: **conoscere l'indirizzo della destinazione non ci aiuta**
  - potremmo, certo, scegliere, fra i nostri vicini quello le cui coordinate sono più prossime a quelle di v
  - ma la casualità dei weak ties potrebbe far sì che, invece, un vicino che, al momento, mi appare peggiore ha un arco random che lo collega direttamente a v…

2) Percorsi brevi facili da trovare

- In figura:
  - s deve inviare un messaggio a d
  - fra i vicini di s il più vicino a d sulla griglia è a
  - perché b sulla griglia è più lontano di a da d
  - e questo è tutto ciò che s sa
  - perciò, s inoltra il messaggio ad a
    - che poi dovrà seguire i nodi della griglia per giungere a d
  - invece, se s avesse inoltrato a b
    - allontanandosi momentaneamente da d
  - avrebbe raggiunto d in 3 passi!



2) Percorsi brevi facili da trovare

- Ora: siamo il nodo u in un grafo di Watts-Strogatz e vogliamo inviare un messaggio ad una certa destinazione v **di cui conosciamo le coordinate**
- allora, facciamo tante copie del messaggio quanti sono i nostri vicini
  - tanto, abbiamo soltanto 8/9 vicini – non andiamo falliti in fotocopie
- e mandiamo una copia a ciascun vicino che, a sua volta, conosce solo i propri vicini… e non può far altro che ripetere il nostro procedimento
  - ora, a parte che non è carino chiedere ai nostri amici di spendere soldi in fotocopie per spedire un nostro messaggio
  - a parte questo, dopo h passi circoleranno nel grafo $\approx 7^h$ copie del messaggio!
- Tecnicamente parlando, per spedire un messaggio da un nodo u a un nodo v, è stato utilizzato un ***flooding***:
  - quando un nodo entra in possesso del messaggio, crea di esso tante copie quanti sono i suoi vicini
  - e invia una copia a ciascun vicino

2) Percorsi brevi facili da trovare

- Il meccanismo del flooding determina un sovraccarico della rete
  - dopo h passi circoleranno nel grafo ≈ 7^h copie del messaggio!
  - Un carico inaccettabile
- Di contro, nell'esperimento di Milgram, ad ogni passo, circolavano tanti messaggi quanti ne aveva creati Milgram
  - i messaggi non aumentavano ad ogni passo
  - per consegnare il messaggio attraverso uno shortest path, non veniva usato il flooding
- Eppure, anche gli individui che parteciparono all'esperimento di Milgram avevano conoscenza solo dei propri vicini e dell'indirizzo del destinatario
  - esattamente le stesse informazioni che hanno i nodi di un grafo di Watts-Strogatz!
- Allora, evidentemente, il modello di Watts-Strogatz non riesce a descrivere qualche caratteristica di una rete sociale che rende possibile la ricerca di Milgram

2) Percorsi brevi facili da trovare

- Un algoritmo di ***ricerca decentralizzata*** (o ***ricerca miope***) è tale che
  - ciascun nodo non conosce della rete altro che i propri vicini (oltre al nodo target della ricerca)
  - e i nodi non comunicano in alcun modo se non nell'invio del messaggio da consegnare
  - Vincoli ancor più restrittivi che in un algoritmo distribuito!

- Gli individui che parteciparono all'esperimento di Milgram avevano conoscenza solo dei propri vicini e dell'indirizzo del destinatario
- In sostanza, gli individui coinvolti da Milgram nel suo esperimento hanno utilizzato un algoritmo di ricerca decentralizzata
  - che, in quel caso, si dimostrava efficiente!

- Osserviamo: probabilmente, ad ogni passo, un individuo inoltrava la copia della lettera in suo possesso a quello fra i suoi contatti che stimava essere più vicino possibile al destinatario
  - dove "più vicino" può riferirsi a metriche diverse: più vicino geogaficamente, o rispetto alla professione, o agli interessi culturali...

2) Percorsi brevi facili da trovare

- Probabilmente, ad ogni passo, ciascun individuo inoltrava la copia della lettera in suo possesso a quello fra i suoi contatti che stimava essere più vicino possibile al destinatario
- Allora, la struttura della rete ha qualche caratteristica che fa sì che “inoltrare al più vicino” funziona bene
  - ossia, in qualche modo, la struttura della rete garantisce che, se ad un passo sono nel nodo u e invio la lettera al nodo v, allora fra gli amici di v c’è (probabilmente) qualcuno ***molto*** più vicino di u alla destinazione
  - ossia, la struttura della rete garantisce che ad ogni passo mi avvicino alla destinazione
  - e che, *probabilmente*, ci sono *<u>tanti passi</u>* in cui la distanza dalla destinazione *<u>diminuisce</u>* ***<u>drasticamente</u>***
- Riconsideriamo ora il modello di Watts-Strogatz:
  - certamente, ciascun nodo ha un vicino sulla griglia più vicino di sé alla destinazione
  - d’altra parte, se seguiamo un percorso costituito di soli archi della griglia, impieghiamo un sacco di passi per giungere a destinazione – $O(\sqrt{n})$ passi
  - allora, se voglio trovare un percorso *molto* breve, devo usare gli archi random

2) Percorsi brevi facili da trovare

- Nel modello di Watts-Strogatz:
  - certamente, ciascun nodo ha un vicino sulla griglia più vicino di sé alla destinazione
  - d'altra parte, se voglio trovare un percorso breve, devo usare gli archi random
- Ma non c'è garanzia che, in un grafo di Watts-Strogatz,
  - usando la regola "invia al tuo vicino che è il più vicino alla destinazione"
  - incontrerò una serie di archi random in modo tale che in un piccolo numero di passi giungerò a destinazione
- In effetti, si può dimostrare che nel modello di Watts-Strogatz la ricerca decentralizzata di un percorso da s a t individua mediamente un percorso molto più lungo di uno shortest path [Kleinberg, 2000]
- Perché nel modello di Watts-Strogatz l'estremo di un arco random uscente da un nodo è scelto *uniformemente a caso* fra tutti gli altri nodi
- gli archi random non tengono conto in alcun modo di quanto sono "vicini" i nodi che congiungono
  - qualunque significato decidiamo di associare a "vicini"
- Detto altrimenti, **gli archi random sono *troppo* random**!

Un modello per la ricerca decentralizzata

- Quindi, vogliamo definire un modello generativo cui corrispondano grafi
  - che contengono molti triangoli,
  - nei quali esistono molti shortest paths fra le coppie di nodi, e
- nei quali **trovare gli shortest paths mediante ricerca decentralizzata sia possibile**
- Dal nostro ragionamento intuitivo, possiamo ben pensare che, per soddisfare l'ultimo punto, è necessario che gli archi random siano scelti in modo da tener conto di quanto sono "vicini" i nodi che congiungono

- Il nostro nuovo modello è ancora basato su un'ossatura deterministica:
  - la stessa griglia arricchita e wrapped del modello di Watts-Strogatz
- e da ogni nodo esce un arco random
- ma ora la probabilità che l'arco random uscente dal nodo u sia (u,v) è inversamente proporzionale alla distanza sulla griglia dei nodi u e v
- ossia, **P((u,v) ∈ E) =** $\frac{1}{Z_u} \frac{1}{d(u,v)^q}$

Un modello per la ricerca decentralizzata

- La probabilità che l'arco random uscente dal nodo u sia (u,v) è inversamente proporzionale alla distanza sulla griglia dei nodi u e v
- ossia, $\mathbf{P((u,v) \in E) = \frac{1}{Z_u} \frac{1}{d(u,v)^q}}$ dove
- d(u,v) indica la lunghezza di uno shortest path fra u e v **sulla griglia**
  - (ossia, un percorso che non contiene gli archi random)
- Z_U è un fattore di normalizzazione:
  - poiché da ogni nodo deve uscire uno e un solo arco random, deve essere $\sum_{v \in V - \{u\}} P((u,v) \in E) = 1$
  - ossia, $\sum_{v \in V - \{u\}} \frac{1}{Z_u} \frac{1}{d(u,v)^q} = 1$ e quindi $\mathbf{Z_U = \sum_{v \in V - \{u\}} \frac{1}{d(u,v)^q}}$
  - Poiché la griglia wrapped è simmetrica, allora Z_U ha lo stesso valore per tutti i nodi e, pertanto, **indicheremo il fattore di normalizzazione, semplicemente, come Z**
- q è un parametro che prende il nome di **esponente di clustering**
  - abbiamo un modello diverso per ogni valore di q: ad esempio, per q = 0 abbiamo il modello di Watts-Strogatz
  - in generale, gli archi random sono "troppo random" quando q è piccolo, "poco random" quando q è grande

Un modello per la ricerca decentralizzata

- **P((u,v) ∈ E) =** $\frac{1}{Z}\ \frac{1}{d(u,v)^q}$
  - con **Z =** $\sum_{v\,\in V\,-\{u\}}\ \frac{1}{d(u,v)^q}$
- q è un parametro che prende il nome di **esponente di clustering**
  - abbiamo un modello diverso per ogni valore di q
  - ad esempio, per q = 0 abbiamo il modello di Watts-Strogatz
  - in generale, gli archi random sono "troppo random" quando q è piccolo, "non abbastanza random" quando q è grande
- Naturalmente, la ricerca decentralizzata funziona meglio con alcune scelte di q e peggio con altre
  - già abbiamo visto, che non funziona bene con q=0
- Quel che ci proponiamo è mostrare che esiste una scelta di q che rende efficiente la ricerca decentralizzata
  - ossia, permette di trovare percorsi la cui lunghezza non è troppo lontana da quella degli shortest paths
- anzi, che ***esiste un valore di q ottimale per la ricerca decentralizzata***

Un modello per la ricerca decentralizzata

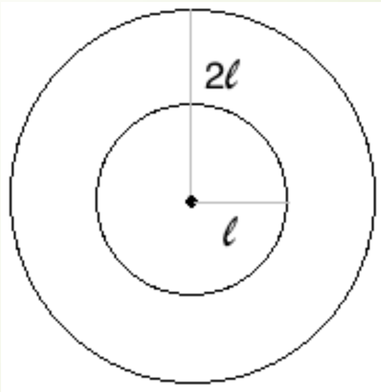
- **P((u,v) ∈ E) =** $\frac{1}{Z}\,\frac{1}{d(u,v)^q}$ con **Z =** $\sum_{v \in V - \{u\}} \frac{1}{d(u,v)^q}$
- Ci proponiamo di mostrare che esiste un valore di q ottimale per la ricerca decentralizzata
- Ebbene, **nel caso in cui la componente deterministica del grafo è una griglia (wrapped) bidimensionale**
  - ossia, nodi sono immersi in una superficie (bidimensionale)
- **allora l'esponente di clustering ottimale è q = 2**
  - quando il numero di nodi è molto, molto grande
- In generale, se la componente deterministica è una griglia (wrapped) d-dimensionale
  - ossia, nodi sono immersi nello spazio $\mathbb{R}^d$
- allora l'esponente di clustering ottimale è q = d
- ed ora vediamo qualche intuizione in supporto di questa affermazione

Un modello per la ricerca decentralizzata

- $\mathbf{P((u,v) \in E) = \frac{1}{Z} \frac{1}{d(u,v)^q}}$ con $\mathbf{Z} = \sum_{v \in V - \{u\}} \frac{1}{d(u,v)^q}$
- Se la componente deterministica del grafo è una griglia bidimensionale (ossia, i nodi giacciono su una superficie), allora l'esponente di clustering ottimale è q = 2
- Ripensiamo l'esperimento di Milgram:
  - se il destinatario della lettera che Milgram consegna al newyorkese Pippo vive all'altro capo del mondo rispetto a Pippo, diciamo a Roma, quartiere Garbatella, Pippo cerco fra i suoi amici qualcuno che viva in Europa
  - diciamo che Pippo un amico che vive in Europa ce l'ha: è Pluto, vive a Mosca
  - Pluto cerca fra i suoi amici qualcuno che viva in Europa occidentale, e così la lettera finisce a Parigi
  - poi a Milano (Italia), Perugia (Italia centrale), Latina (Lazio), Roma Centocelle, Roma Garbatella (finalmente a destinazione)
- Ossia, il tragitto della lettera segue, grosso modo, uno schema a "scale di risoluzione"
  - prima viaggia da un continente all'altro, poi all'interno del continente, poi all'interno della nazione, della regione, della città
- Se immaginiamo il viaggio su una mappa elettronica, ad ogni passo aumenta la risoluzione di ciò che viene visualizzato sulla mappa

Un modello per la ricerca decentralizzata

- $\mathbf{P((u,v) \in E) = \frac{1}{Z} \frac{1}{d(u,v)^q}}$ con $\mathbf{Z} = \sum_{v \in V - \{u\}} \frac{1}{d(u,v)^q}$
- Se la componente deterministica del grafo è una griglia bidimensionale (ossia, nodi giacciono su una superficie), allora l'esponente di clustering ottimale è q = 2
- Ora proviamo a formalizzare quel che abbiamo osservato e vediamo a cosa ci porta
- Fissiamo un nodo u e
  - partizioniamo i nodi rimanenti per blocchi definiti in base alla distanza da u: i nodi a distanza da u compresa fra 2 e 4, quelli a distanza fra 4 e 8, … , quelli a distanza fra 2^h e 2^{h+1}, ...
  - il numero di nodi nel blocco $2^{h+1} - 2^h$ è $\approx \pi(2^{h+1})^2 - \pi 2^{2h} = \pi 2^{2h} (4-1) = 3 \pi 2^{2h}$, ossia è **proporzionale a 2^{2h}**
  - scelto v nel blocco $2^{h+1} - 2^h$, la probabilità che l'arco random uscente da u sia (u,v) **è proporzionale a $1/2^{2h}$** (*perché q=2 !*)
  - **allora la probabilità che l'arco random uscente da u cada nel blocco $2^{h+1} - 2^h$ è indipendente da h**
  - ossia, è indipendente da quale blocco si stia considerando
  - ossia, la probabilità di raggiungere un nodo a distanza 2, o 4, .. , o 64, o 1024 (ecc. ecc.) è la stessa!

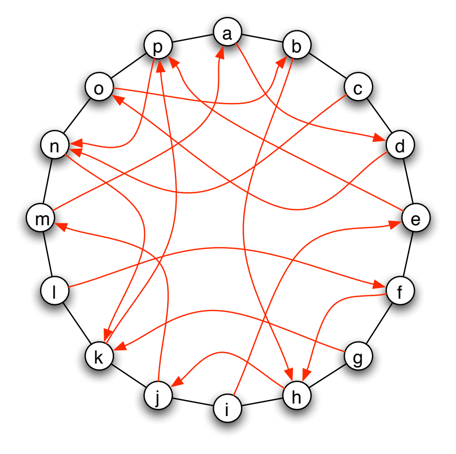


Un modello per la ricerca decentralizzata

- $P((u,v) \in E) = \frac{1}{Z} \frac{1}{d(u,v)^q}$ con $Z = \sum_{v \in V - \{u\}} \frac{1}{d(u,v)^q}$
- Se la componente deterministica del grafo è una griglia bidimensionale (ossia, nodi giacciono su una superficie), allora l'esponente di clustering ottimale è q = 2
- Intuitivamente, perché *la probabilità che l'arco random uscente da u cada in un certo blocco è la stessa per tutti i blocchi*
- Ossia, ***i weak ties sono distribuiti uniformemente su tutte le scale di risoluzione***
- e questo fa sì che, anche se per un certo numero di passi occorre utilizzare gli archi della griglia,
  - perché gli archi random che si incontrano fanno allontanare dall'obiettivo
  - e, dunque, ad ognuno di questi passi ci si avvicina solo di un'inezia all'obiettivo
- non occorreranno molti passi prima di arrivare ad un nodo il cui arco random diminuisce ***drasticamente*** la distanza dall'obiettivo
  - la diminuisce, cioè, di un ordine di grandezza!
- E ora, dopo l'intuizione, l'ottimalità di q = d nel caso d -dimensionale non ci resta che dimostrarla formalmente

Prestazioni della ricerca miope nel modello

- Analizziamo formalmente le prestazioni dell'algoritmo di ricerca decentralizzata applicata al modello generativo che abbiamo introdotto nel solo caso d =1
  - perché, naturalmente, l'analisi risulta più semplice
  - anche se la generalizzazione ad altre dimensioni è, sostanzialmente, basata sugli stessi argomenti
- Ossia, analizziamo il caso in cui i nodi sono in uno spazio unidimensionale
- ossia, sono disposti su un anello
- al quale sono aggiunti gli archi random, in accordo al modello che abbiamo descritto
- **con coefficiente di clustering q = 1**



(b) *A ring augmented with random long-range links.*

Prestazioni della ricerca miope nel modello

- Più precisamente, consideriamo un grafo G tale che:
- i nodi sono sono disposti su un anello,
  - ossia, $V = [n]$
  - e $\{ (i, i+1): 1 \leq i < n\} \cup \{(n,1)\} \subseteq E$
- al quale sono aggiunti gli archi random: per $u,v \in V$, $\mathbf{P((u,v) \in E) = \frac{1}{Z} \frac{1}{d(u,v)}}$
  - perché **q = 1**
  - con $Z = \sum_{v \in V - \{u\}} \frac{1}{d(u,v)}$
- Osserviamo che, nel nostro caso (nodi su un anello e q = 1), per ogni $u \in V$,

$$\mathbf{Z} = \sum_{v \in V - \{u\}} \frac{1}{d(u,v)} = 2 \sum_{1 \leq h \leq \frac{n}{2}} \frac{1}{h}$$

perché in un anello u ha due vicini a distanza 1, due vicini a distanza 3, e così via

$$\leq 2 \sum_{h=1}^{\infty} \frac{1}{h} = \mathbf{2 \ln n}$$

Prestazioni della ricerca miope nel modello

- Sia G tale che: i nodi sono sono disposti su un anello, al quale sono aggiunti gli archi random:
  - per u,v ∈ V, P((u,v) ∈ E) = $\frac{1}{Z} \frac{1}{d(u,v)}$ con $Z = \sum_{v \in V - \{u\}} \frac{1}{d(u,v)}$
  - ove **$Z \leq 2 \ln n$**
- Scegliamo uniformemente a caso due nodi s e t in G
- ed utilizziamo l'algoritmo di ricerca decentralizzata per calcolare un percorso da s a t
- Indichiamo con X la variabile aleatoria che denota la lunghezza di tale percorso
- Allora,

Teorema: $E[X] \in O(\ln^2 n)$

Prestazioni della ricerca miope nel modello

- **Teorema: E[X] ∈ O(ln^2 n)**
- per fissare le idee, visualizziamo l'esecuzione dell'algoritmo di ricerca decentralizzata che costruisce in G un percorso da s a t come segue:
  - ricordiamo che, inizialmente, s possiede una copia della lettera
  - che al passo 1 trasmette ad un suo vicino (a quello più prossimo alla destinazione),
  - che, a sua volta, al passo 2, trasmette al suo vicino più prossimo alla destinazione, e così via fino a quando la lettera raggiunge t
- suddividiamo in fasi il processo di trasmissione della lettera da un nodo all'altro:
  - durante la **fase j** la lettera è in possesso di un nodo u tale che

$$\frac{d(s,t)}{2^{j+1}} < d(u,t) \leq \frac{d(s,t)}{2^j}$$

  - cosicché, il processo inizia con la fase 0
- Poiché d(s,t) $\leq \frac{n}{2}$ e ad ogni fase si dimezza la distanza fra il nodo che possiede la lettera e t, allora **il numero di fasi è $\leq \log_2 n$**

Prestazioni della ricerca miope nel modello

- **Teorema: E[X] ∈ O(ln^2 n)**
- durante la fase j la lettera è in possesso di un nodo u tale che $\frac{d(s,t)}{2^{j+1}}$ < d(u,t) ≤ $\frac{d(s,t)}{2^j}$
- il numero di fasi è ≤ $\log_2 n$
- Indichiamo con X_j la durata della fase j
  - ossia, X_j è il numero di nodi che entrano in possesso della lettera durante la fase j
- Allora, X = $\sum_{1 \leq j \leq \log_2 n} X_j$ e **E[X] = $\sum_{1 \leq j \leq \log_2 n} E[X_j]$**
- Per dimostrare il teorema è, quindi, sufficiente dimostrare che,

  per ogni j, E[X_j] ∈ O($\ln n$)

Prestazioni della ricerca miope nel modello

- **Teorema: $E[X] \in O(\ln^2 n)$**
  - X_j è la durata della fase j , $X = \sum_{1 \leq j \leq \log_2 n} X_j$ e $E[X] = \sum_{1 \leq j \leq \log_2 n} E[X_j]$
- **Dimostriamo che, per ogni j, $E[X_j] \in O(\ln n)$**
  - Supponiamo di trovarci nel nodo v durante la fase j: allora, $\frac{d(s,t)}{2^{j+1}} < d(v,t) \leq \frac{d(s,t)}{2^j}$
  - La fase j termina sicuramente se esiste un nodo z tale che

$$(v,z) \in E \text{ e } d(z,t) \leq \frac{d(v,t)}{2}$$

    - perché $d(z,t) \leq \frac{d(v,t)}{2} \leq \frac{1}{2}\frac{d(s,t)}{2^j}$
  - quindi, **P(la fase j termina)** = $P(\exists\, z \in V\colon\ (v,z) \in E \text{ e } d(z,t) \leq \frac{d(s,t)}{2^{j+1}})$

$$\geq \mathbf{P(\exists\, z \in V\colon\ (v,z) \in E \text{ e } d(z,t) \leq \frac{d(v,t)}{2}\ )}$$

    - $\geq$: perché potrebbe essere $d(v,t) = \frac{d(s,t)}{2^j}$
      - in tal caso esisterebbe un arco (v,z) *dell'anello,* ossia tale che $d(z,t) = d(v,t) - 1 < \frac{d(s,t)}{2^j}$, e tale arco farebbe certamente terminare la fase j
      - e tuttavia sarebbe $d(z,t) > \frac{d(v,t)}{2}$

Prestazioni della ricerca miope nel modello

- **Teorema: $E[X] \in O(\ln^2 n)$**
  - X_j è la durata della fase j , $X = \sum_{1 \leq j \leq \log_2 n} X_j$ e $E[X] = \sum_{1 \leq j \leq \log_2 n} E[X_j]$
- **Dimostriamo che, per ogni j, $E[X_j] \in O(\ln n)$**
  - Supponiamo di trovarci nel nodo v durante la fase j: allora, $\frac{d(s,t)}{2^{j+1}} < d(v,t) \leq \frac{d(s,t)}{2^j}$
  - **P(la fase j termina) $\geq$ P($\exists$ z $\in$ V: (v,z) $\in$ E e d(z,t) $\leq \frac{d(v,t)}{2}$)**
  - Indichiamo con $\mathbf{I}$ l'insieme dei nodi che distano da t non più della metà di quanto v dista da t: **$\mathbf{I}$= {u $\in$ V: d(u,t) $\leq \frac{d(v,t)}{2}$ }**
  - Allora, **P($\exists$ z $\in$ V: (v,z) $\in$ E e d(z,t) $\leq \frac{d(v,t)}{2}$)**

= P($\exists$ z $\in \mathbf{I}$: (v,z) $\in$ E)

= P($\exists$ z $\in \mathbf{I}$: (v,z) è un arco dell'anello oppure è un arco random)

= $\sum_{z \in \mathbf{I}}$ P((v,z) è un arco dell'anello oppure (v,z) è un arco random)

$\geq \sum_{z \in \mathbf{I}}$ P((v,z) è un arco random)

Prestazioni della ricerca miope nel modello

- **Teorema: $E[X] \in O(\ln^2 n)$**
  - **$I = \{u \in V:\ d(u,t) \leq \frac{d(v,t)}{2}\}$**
  - Allora, **$P(\exists\, z \in V:\ (v,z) \in E \text{ e } d(z,t) \leq \frac{d(v,t)}{2}) \geq \sum_{z \in I} P(\,(v,z) \text{ è un arco random})$**
  - sia $z \in I$: allora

$$d(v,z) \leq d(v,t) + d(t,z) = d(v,t) + d(z,t) \leq d(v,t) + \frac{d(v,t)}{2} = \frac{3\, d(v,t)}{2}$$

  - allora, per ogni $z \in I$:

$$P(\,(v,z) \text{ è un arco random}) = \frac{1}{Z}\ \frac{1}{d(v,z)} \geq \frac{1}{Z}\ \frac{2}{3\, d(v,t)}$$

$$\geq \frac{1}{2\ \ln n}\ \frac{2}{3\, d(v,t)} \qquad \text{perché } Z \leq 2 \ln n$$

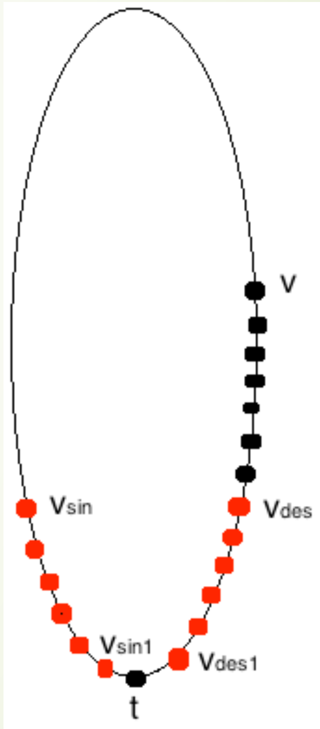
$$= \frac{1}{3\, d(v,t)\ \ln n}$$

  - allora, **$P(\exists\, z \in V:\ (v,z) \in E \text{ e } d(z,t) \leq \frac{d(v,t)}{2}) \geq \sum_{z \in I} \frac{1}{3\, d(v,t)\ \ln n} = \frac{1}{3\, d(v,t)\ \ln n}\ |I|$**

Prestazioni della ricerca miope nel modello

- **Teorema: E[X] $\in$ O(ln^2 n)**
  - **P(la fase j termina)** $\geq$ **P($\exists$ z $\in$ V: (v,z) $\in$ E e d(z,t) $\leq \frac{d(v,t)}{2}$)** $\geq \frac{1}{3\, d(v,t)\ \ln n}$ $|\mathbf{I}|$
  - resta da valutare $|\mathbf{I}|$ = | **{u $\in$ V: d(u,t) $\leq \frac{d(v,t)}{2}$ }** |
    - siano v_{sin} e v_{des} i due nodi in **I** a distanza massima da t
    - allora, d(v_{sin} , t) = d(v_{des} , t) = $\left\lfloor \frac{d(v,t)}{2} \right\rfloor$
      - parte intera inferiore, perché $\frac{d(v,t)}{2}$ potrebbe non essere intero
    - e siano v_{sin1} e v_{des1} i due nodi adiacenti a t
    - allora **I** contiene: t, poi i $\left\lfloor \frac{d(v,t)}{2} \right\rfloor$ nodi da v_{sin} a v_{sin1},

      poi i $\left\lfloor \frac{d(v,t)}{2} \right\rfloor$ nodi da v_{des} a v_{des1},
    - Allora: $|\mathbf{I}|$ = 1 + $\left\lfloor \frac{d(v,t)}{2} \right\rfloor$ + $\left\lfloor \frac{d(v,t)}{2} \right\rfloor \geq 1 + \frac{d(v,t)-1}{2} + \frac{d(v,t)-1}{2}$ **= d(v,t)**
  - In conclusione, **P(la fase j termina)** $\geq \frac{1}{3\, d(v,t)\ \ln n}$ **d(v,t)**



Prestazioni della ricerca miope nel modello

- **Teorema: $E[X] \in O(\ln^2 n)$**
  - Indichiamo con X_j la durata della fase j
  - Allora, $X = \sum_{1 \leq j \leq \log_2 n} X_j$ e $E[X] = \sum_{1 \leq j \leq \log_2 n} E[X_j]$
- **Dimostriamo che, per ogni j, $E[X_j] \in O(\ln n)$**
  - P(la fase j termina) $\geq \frac{1}{3\,d(v,t)\,\ln n}\,d(v,t) = \frac{1}{3\,\ln n}$
  - Allora, P(la fase j <u>non</u> termina) $\leq 1 - \frac{1}{3\,\ln n}$
  - Allora, **$P(X_j \geq h)$ = P(la fase j <u>non</u> termina per h passi) $\leq \left[1 - \frac{1}{3\,\ln n}\right]^h$**
  - Non resta che calcolare $E[X_j]$:

    $E[X_j] = 1 \cdot P(X_j = 1) + 2 \cdot P(X_j = 2) + 3 \cdot P(X_j = 3) + \ldots + \frac{n}{2} \cdot P(X_j = \frac{n}{2})$ perché $d(s,t) \leq \frac{n}{2}$

Prestazioni della ricerca miope nel modello

- **Teorema: E[X] ∈ O($\ln^2 n$)**
  - Indichiamo con X_j la durata della fase j
  - Allora, $X = \sum_{1 \le j \le \log_2 n} X_j$ e $E[X] = \sum_{1 \le j \le \log_2 n} E[X_j]$
- **Dimostriamo che, per ogni j, $E[X_j] \in O(\ln n)$**
  - **$P(X_j \geq h) \leq \left[1 - \frac{1}{3 \ln n}\right]^h$**
  - $E[X_j] = 1 \cdot P(X_j = 1) + 2 \cdot P(X_j = 2) + 3 \cdot P(X_j = 3) + 4 \cdot P(X_j = 4) + \ldots + \frac{n}{2} \cdot P(X_j = \frac{n}{2})$

$$= P(X_j \geq 1) + 1 \cdot P(X_j = 2) + 2 \cdot P(X_j = 3) + 3 \cdot P(X_j = 4) + \ldots + \left(\frac{n}{2} - 1\right) \cdot P(X_j = \frac{n}{2})$$
$$= P(X_j \geq 1) + P(X_j \geq 2) + 1 \cdot P(X_j = 3) + 2 \cdot P(X_j = 4) + \ldots + \left(\frac{n}{2} - 2\right) \cdot P(X_j = \frac{n}{2})$$
$$= P(X_j \geq 1) + P(X_j \geq 2) + P(X_j \geq 3) + 1 \cdot P(X_j = 4) + \ldots + \left(\frac{n}{2} - 3\right) \cdot P(X_j = \frac{n}{2})$$
$$= \ldots$$
$$= P(X_j \geq 1) + P(X_j \geq 2) + P(X_j \geq 3) + P(X_j \geq 4) + \ldots + P(X_j \geq \frac{n}{2})$$
$$= \sum_{1 \le h \le \frac{n}{2}} P(X_j \geq h)$$

Prestazioni della ricerca miope nel modello

- **Teorema: $E[X] \in O(\ln^2 n)$**
  - Indichiamo con X_j la durata della fase j
  - Allora, $X = \sum_{1 \le j \le \log_2 n} X_j$ e $E[X] = \sum_{1 \le j \le \log_2 n} E[X_j]$
- **Dimostriamo che, per ogni j, $E[X_j] \in O(\ln n)$**
  - $P(X_j \ge h) \le \left[1 - \frac{1}{3 \ln n}\right]^h$
  - $E[X_j] = \sum_{1 \le h \le \frac{n}{2}} P(X_j \ge h) \le \sum_{1 \le h \le \frac{n}{2}} \left[1 - \frac{1}{3 \ln n}\right]^h$

    $\le \sum_{h \ge 0} \left[1 - \frac{1}{3 \ln n}\right]^h = \frac{1}{1 - \left[1 - \frac{1}{3 \ln n}\right]} = 3 \ln n$
- **Terminiamo la dimostrazione**
  - $E[X] = \sum_{1 \le j \le \log_2 n} E[X_j] \le \sum_{1 \le j \le \log_2 n} 3 \ln n = \log_2 n \cdot 3 \ln n$

    $= \log_2 e \cdot \ln n \cdot 3 \ln n \in O(\ln^2 n)$

QED

Perché q = d funziona bene in $\mathbb{R}^d$

- Gli "ingredienti" che permettono di dimostrare che la ricerca decentralizzata si comporta "bene" nell'anello (q = d =1) sono:
  - 1) **fissato d, il numero di nodi a distanza (al più) d dalla destinazione t è, all'incirca, d**
  - 2) il fattore di normalizzazione è Z ≤ 2 ln n
- Questi due "ingredienti" permettono di dimostrare che, trovandoci in un nodo v a distanza d da t, la probabilità che v abbia un vicino z a distanza $\leq \frac{d}{2}$ da t è proporzionale a $\frac{1}{\ln n}$ ***indipendentemente da* d**
  - Infatti, detto **I= {u ∈ V: d(u,t) ≤ $\frac{d(v,t)}{2}$ }** , **|I|= d(v,t)**
  - **P(∃ z ∈ V: (v,z) ∈ E e d(z,t) ≤ $\frac{d(v,t)}{2}$) ≥ $\frac{1}{3\, d(v,t)\ \ln n}$ |I| = $\frac{1}{3\ \ln n}$**
- considerazioni analoghe possono essere fatte nel **caso bidimensionale con q = 2**:
  - 1) ora , **|I|= α d(v,t)^2** , per qualche costante α (area di un quadrato di lato ≈ d(v,t))
  - 2) ora Z = $\sum_{v \in V-\{u\}} \frac{1}{d(u,v)^2}$ e si dimostra che Z ≤ β ln n, per qualche costante β
  - allora, **P(∃ z ∈ V: (v,z) ∈ E e d(z,t) ≤ $\frac{d(v,t)}{2}$)** $\geq \sum_{z \in I} \frac{1}{Z\, d(v,z)^2} \geq \frac{4}{Z\, 9\, d(v,t)^2}\ |I| \geq \boldsymbol{\frac{4\alpha}{9\, \beta\, \ln n}}$
- allora, anche nel caso q = d = 2 la probabilità che v abbia un vicino a distanza $\leq \frac{d}{2}$ da t è proporzionale a $\frac{1}{\ln n}$ indipendentemente da d
- Allo stesso modo, considerazioni analoghe valgono anche per d > 2

Perché q ≠ d funziona male in $\mathbb{R}^d$

- Cerchiamo, ora, di capire perché il nostro modello nel caso d = 1 e q = 0 mal si presta alla ricerca decentralizzata
  - ossia: quando si esegue la ricerca decentralizzata di un percorso in un anello con archi random generati in accordo al modello con q = 0, si trova mediamente un percorso la cui lunghezza è parecchio elevata

- La probabilità che l'arco random uscente dal nodo u sia (u,v) è, in questo caso, **P((u,v) ∈ E) =** $\frac{1}{Z}\frac{1}{d(u,v)^0} = \frac{1}{Z}$ dove
  - poiché deve essere $\sum_{v \in V - \{u\}}$ P((u,v) ∈ E) = 1, allora
  - ossia, $\sum_{v \in V - \{u\}} \frac{1}{Z} = 1$ e quindi **Z =** $\frac{1}{n-1}$
- da cui la probabilità che l'arco random uscente dal nodo u sia (u,v) è

$$P(\,(u,v) \in E) = \frac{1}{n-1}$$

Perché q ≠ d funziona male in $\mathbb{R}^d$

- Cerchiamo, ora, di capire perché il nostro modello nel caso d = 1 e q = 0 mal si presta alla ricerca decentralizzata
  - ossia: quando si esegue la ricerca decentralizzata di un percorso in un anello con archi random generati in accordo al modello con q=0, si trova mediamente un percorso la cui lunghezza è parecchio elevata
- La probabilità che l'arco random uscente dal nodo u sia (u,v) è ora

$$\mathbf{P((u,v) \in E) = \frac{1}{n-1}}$$

- Come abbiamo già osservato, nel caso "anello & q =1" è "facile" entrare in regioni del grafo contenenti nodi sempre più vicini a t – gli insiemi $\mathbf{I}$
  - più precisamente, la probabilità di *dimezzare* la nostra distanza da t è la stessa sia quando siamo molto lontani da t che quando siamo molto vicini a t
- Mostriamo ora (molto informalmente, tanto per farci un'idea) che nel caso "anello & q = 0" è "difficile" entrare nell'insieme

$$R = \{ u \in [n]: d(u,t) \leq \sqrt{n} \}$$

Perché q ≠ d funziona male in $\mathbb{R}^d$

- Mostriamo ora (molto informalmente, tanto per farci un'idea) che nel caso "anello & q=0" è "difficile" entrare nell'insieme R = { u ∈ [n]: d(u,t) ≤ $\sqrt{n}$ }
  - scegliamo t, e poi scegliamo s ∉ R
  - allora ∀ u ∈ R e ∀ v ∈ [n]-R, P((v,u) ∈ E) > $\frac{1}{n}$
    - ">" sia perché $\frac{1}{n-1} > \frac{1}{n}$, sia perché $\frac{1}{n-1}$ è la probabilità di esistenza di un arco random, ma u e v potrebbero anche essere vicini lungo l'anello
  - allora, ∀ v ∉ R, P(∃ u ∈ R: (v,u) ∈ E) > $\frac{|R|}{n} = \frac{2\sqrt{n}}{n} = \frac{2}{\sqrt{n}}$
  - allora, detta Y la variabile aleatoria che rappresenta il numero di passi per raggiungere da s un nodo in R, P(Y ≥ h) < $\left(1 - \frac{2}{\sqrt{n}}\right)^h$
  - e E[Y] = $\sum_{h \geq 0}$ P(Y ≥ h) ≤ $\sum_{h \geq 0} \left(1 - \frac{2}{\sqrt{n}}\right)^h = \frac{1}{1 - \left(1 - \frac{2}{\sqrt{n}}\right)} = \frac{\sqrt{n}}{2}$
  - che è **esponenzialmente** più grande di E[X] ∈ O($\ln^2 n$) del caso q = 1 !

Perché q ≠ d funziona male in $\mathbb{R}^d$

- Abbiamo visto (informalmente) che nel caso "anello & q = 0" entrare nella regione R = { u ∈ [n]: d(u,t) ≤ $\sqrt{n}$ } richiede, mediamente, $\frac{\sqrt{n}}{2}$ passi
- A questo punto, vediamo in quanti passi, mediamente, entriamo nella regione R_2 = { u ∈ [n]: d(u,t) ≤ $\frac{\sqrt{n}}{2}$} :
  - analogamente a quanto fatto nel caso di R, ∀ v ∈ R-R_2,
    $$P(\exists\, u \in R_2: (v,u) \in E) \geq \frac{|R_2|}{n} = \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}}$$
  - e, quindi, $P(Y_2 \geq h) \leq \left(1 - \frac{1}{\sqrt{n}}\right)^h$ - indicando con Y_2 il numero di passi per entrare in R_2
  - e $E[Y_2] = \sum_{h \geq 0} P(Y_2 \geq h) \leq \sum_{h \geq 0} \left(1 - \frac{1}{\sqrt{n}}\right)^h = \frac{1}{1-\left(1 - \frac{1}{\sqrt{n}}\right)} = \sqrt{n}$
- Questo significa che, una volta entrati nella regione R, per entrare in R_2 utilizzare gli archi random è mediamente equivalente a muoversi lungo gli archi dell'anello!
- Ossia, *quando si raggiunge una distanza dall'obiettivo dell'ordine di $\sqrt{n}$, gli archi random non sembrano giocare più alcun ruolo*
- e la ragione di ciò è il fatto che gli archi random sono troppo random
  - come avevamo già osservato per il modello di Watts-Strogatz

Perché q ≠ d funziona male in $\mathbb{R}^d$

- In effetti, quel che abbiamo visto per il caso "anello & q = 0" può essere generalizzato a tutti i casi "anello & 0 < q < 1"
  - in tutti questi casi, gli archi random sono *troppo* random…
- Di contro nel caso "anello & q > 1" gli archi random sono troppo corti
  - e, quindi, è "difficile" imbattersi in archi random che coprano grandi distanze
  - e che permettano di avvicinarsi alla destinazione in pochi passi
- di conseguenza, la ricerca decentralizzata nel caso "anello & q > 1", riesce a fare poco meglio che utilizzare soltanto archi dell'anello per raggiungere la destinazione
- In effetti, si può dimostrare il seguente

Teorema: comunque si scelga q ≠ 1, esistono due costanti positive a_q e c_q tali che,
detta X la variabile aleatoria che esprime la lunghezza del percorso
trovato dall'algoritmo di ricerca decentralizzata
in un anello di n nodi cui sono aggiunti archi random
in accordo al modello che abbiamo descritto,

$$E[X] \geq a_q\, n^{c_q}$$