Now that we're finished with heavy-tails, we've gone through in some detail issues related to

- connectivity – giant component
- heavy-tailed degree distributions

But we haven't talked much about clustering or small diameters.
→ that's what we'll do this class

Each of these alone is not too surprising:

<u>clustering</u> → comes from correlations in edges (two friends of mine are likely to be friends)

<u>small diameters</u> → if I have 100 friends, each of whom has 100 friends, and so on... we get

$\Rightarrow \approx \log n$ diameter

(*) assuming the friends don't overlap too much, e.g. the friends are independently chosen.

But, if we think about (*) in the context of a highly clustered graph, then the combination of small diameter & high clustering becomes <u>surprising!</u>

Graphs that are highly clustered & still

have small diameters are often termed "small world" graphs.

Today's goal:

what causes these short paths to exist?

→ we'll start by looking more in depth at small world experiments such as Milgrams.

→ then we'll look at models that provide insight into the phenomenon.

---

1) Small world experiments: then & now

See ppt

---

2) Modelling the small world.

Now that we've learned a bit more about "small world" properties, I think there are 2 main scientific questions that emerge:

1) Why do such short paths exist?

2) How can people find them without global knowledge?

→ We'll try to answer these by studying simple models....

Additionally, the engineering/commercial question

that Facebook, etc want to answer is:

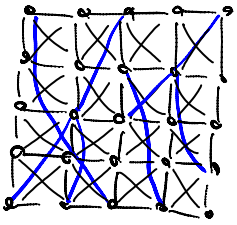
How can we use this to make money?

→ I don't think there's a good answer to this one yet....

Small world models

Q: What is a model where we have lots of triangles (high clustering) but still have a low diameter?

A: One answer (which we'll study today) was introduced by Watts & Strogatz (1998)



A d-dimensional lattice with every node having

- local connections to all lattice pts w/in distance 2.
- long-range connections to non-adjacent lattice points. Each edge is "rewired" w/ prob p to connect to a random other node.

← Creates "Shortcuts"

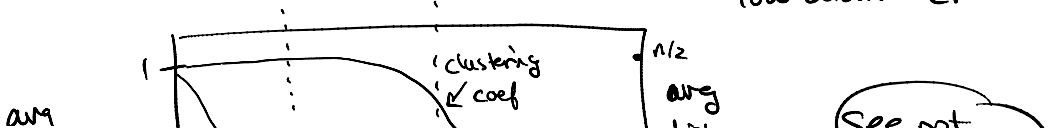
ex: In 1d we have:

See ppt

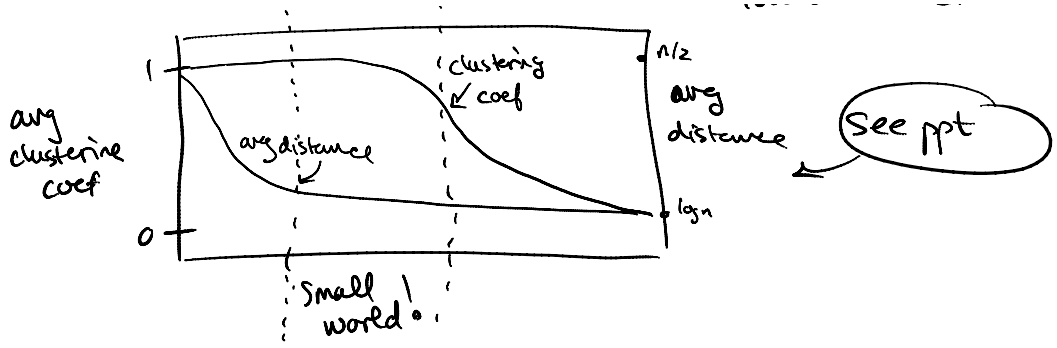
☆ the rewiring allows a transition from high clustering & high diameter

↓

low clustering & low diameter



See ppt

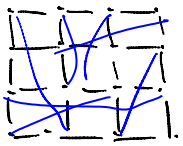


So: One possible explanation for the small world phenomena is a mixture between "local" connections (which give high clustering) & random "long-range" connections (which ensure small diameter)

But: We haven't addressed whether these short paths can be found w/o global information yet.

To start to address this issue, we'll look at a slight variation of the above model.

This is from Kleinberg (2001).



Again we have a d-dim lattice, but this time the local connections are only to nodes w/in distance 1, and there is exactly 1 long distance edge, with it's end point chosen randomly. (For now uniformly.)

→ So, this has the same local/global structure, but now doesn't have triangles.

It's not too hard to show that the diameter is

It's not too hard to show that the diameter is $O(\log n)$, but what we want to know is:

* Can myopic distributed agents find the short paths?

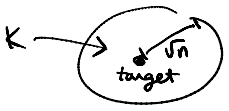
(specifically, consider agents/nodes that
1) know the target destination coordinates.
2) know the coordinates of their neighbors.
3) Forward the "message" to the neighbor closest to the destination.
→ we call these myopic, greedy agents)

Q: Will myopic, greedy agents find short paths in our small world model?

A: No!

We'll consider the 1-d case and show that they will find $\Theta(\sqrt{n})$-length paths.

To see this, consider $K = \{$nodes w/in $\sqrt{n}$ of the target$\}$



As $n \to \infty$, with prob 1 we start outside of K.

Since long range links are chosen uniformly

$Pr(\text{edge connects to node in } K) = \frac{2\sqrt{n}}{n} = \frac{2}{\sqrt{n}}$

$\Rightarrow$ it will take $\frac{\sqrt{n}}{2}$ steps (in expectation to find a node w/ a long range contact in K.

& this is a Geometric distribution, so it is highly concentrated around its mean

$\Rightarrow O(\sqrt{n})$ steps w.p.1.

So, to get to that target, the message must either (i) pass through $K \rightarrow \sqrt{n}$ steps
(ii) have a long-range neighbor in K
$O(\sqrt{n})$ steps w.p.1.

$\Rightarrow O(\sqrt{n})$-steps to get to target.
... which is exponentially larger than the $O(\log n)$ diameter

□

So, we've seen that our small world model doesn't capture the fact that distributed agents can find short paths.

Q: How can we "fix" the model?

A: long range neighbors should be chosen proportionally to their distance.

Change to model:
long-range link chooses neighbors w/prob proportional to $\left(\frac{1}{dist}\right)^q$

see ppt

Q: For which q do you expect distributed agents to be able to find short paths?

A: It turns out that it only works for $q = d$. For all other d, the path length is $O(n^\alpha)$ in expectation, for some α.
↑ the dimension of the lattice
↑ exp larger than the diameter.

Q: Does anyone see why this might be?

Q: Does anyone see why this might be?

A: The proof will provide us the insight... but we can already see that small q are too random to be useful and large q are not random enough since long range links will then be too short

See ppt

What's special about $q=d$ is not obvious, but we'll see it in the proof below

Now on to the main result:

Claim: Greedy, myopic agents will find paths of length $O((\log n)^2)$ in expectation when $q=d$

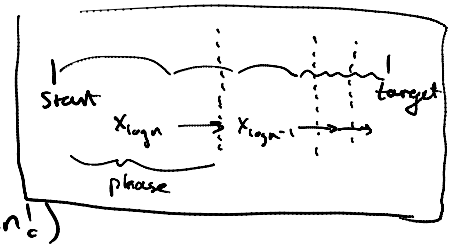
↑ nearly the same as the diameter

pf: We'll prove it for the 1-d case, higher dimensions are essentially the same.

Proof Plan: Break the steps into phases

X_i = steps taken to get from distance 2^{i+1} to 2^i

and $X = \sum X_i = X_1 + \dots + X_{\log n}$ is the total path length.



(Linearity of expectation again!)

with $E[X] = \sum E[X_i]$.

So, there are $\log n$ phases.

The key argument is to show each phase has expected length $O(\log n)$, which gives up $O((\log n)^2)$ overall.

Step 1: Before we can do anything though, we first need to understand the model a little better: we know long-range links happen prop to $\frac{1}{\text{dist}}$, but we need to figure out the normalizing constant, Z.

To do this, let's count the # of nodes at each distance:

there are exactly two at each distance.

So, we know that

$$2Z\left(1+\tfrac{1}{2}+\tfrac{1}{3}+\tfrac{1}{4}+\cdots+\tfrac{1}{n/2}\right) = 1$$

since we have a probability distribution.

$$\Rightarrow \frac{1}{2Z} = 1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n/2}$$

$\approx \log n/2$

$$\leq 1+\int_1^{n/2}\frac{1}{k}\,dk$$

$$= 1+\log_e(n/2)$$

$$\leq 1+\log_2(n)-\log_2(2)$$

$$\leq \log_2(n)$$

$$\Rightarrow \text{prob}(v \text{ links to } w) \geq \frac{1}{2d(v,w)\log n}$$

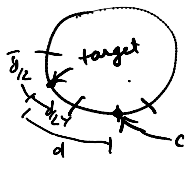
Step 2: Now we'll calculate the time spent in one phase of the search.

→ let's look at phase j so target t is distance d away and $d \in (2^j, 2^{j+1}]$

we'll bound

Pr(long range link of node in phase j enters into Phase $j-1$)

$$\geq \sum_{\substack{\text{nodes w/in} \\ d/2 \text{ of target}}} \Pr(\text{edge exists})$$



$$\geq \left(\frac{d}{2}\cdot 2\right)\left(\frac{1}{2\log n}\cdot\frac{1}{3d/2}\right)$$

↑ # of nodes w/in $d/2$

↑ max distance to node w/in

of target d/2 of target

$$= \frac{d}{\log n} \cdot \frac{1}{3d} = \frac{1}{3}\frac{1}{\log n}$$

$$\Rightarrow E[X_j] = E\left[\begin{array}{c}\#\text{ of steps before a long} \\ \text{range link goes to w/in } d/2 \\ \text{of target}\end{array}\right] \leq 3\log n$$

$$\Rightarrow E[X] \leq (\log n)(3\log n)$$
$$= 3(\log n)^2$$

so greedy, myopic agents find paths only slightly longer than the diameter (which is $\log n$).

□

Now that we've seen the proof,

Q: Why was $k=1$ important?

A: It meant that the long range links were equally distributed across scales. i.e. regardless of the phase, it was equally likely to find a long range link that cut our distance in half.

⇒ This would not have been the case for any other k...

$$E[X_i] \approx \frac{d^i}{k} \cdot \frac{2}{3d} = \Theta(d^{i-1})$$

⇒ some phase would take poly(d) time ... which would be bad.

Summary:

We've seen that short paths in small world graphs come from

correlated local connections combined with random long range connections, but that the long-range connections should be "distance-dependent" for myopic agents to be able to find the short paths.

That's where we'll leave it, but of course there are many other things we could study...

→ Our model didn't have heavy-tails!
→ is "searchability" really so fragile?
→ what if edges change over time?

these turn out to be related...

(see paper on website)

Thanks!