Random models of networks: Erdos Renyi random graphs

(picture from Wikipedia http://en.wikipedia.org/wiki/Erd%C5%91s%E2%80%93R%C3%A9nyi_model)
Preferential attachment models of network formation

Consider the following process to generate a network (e.g. a web graph):

1. Order all of the N pages 1, 2, 3, ..., N and repeat the following process for each page $j$:
2. Use the following rule to generate a link to another page:
   a. With probability $p$, link to a random page $i < j$
   b. Otherwise, choose a random page $i$ and link to the page $i$ links to
• Social and information networks often follow **power laws**, meaning that a few nodes have **many** of the edges, and many nodes have **a few** edges.

- e.g. web graph (Broder et al.)
- e.g. power grid (Barabasi-Albert)
- e.g. Flickr (Leskovec)
We defined the concept of “strong” and “weak” ties (which roughly correspond to notions of “friends” and “acquaintances”)

3. Strong triadic closure property

If (a,b) and (b,c) are connected by strong ties, there must be at least a weak tie between a and c.
How can we **characterize, model, and reason about** the structure of social networks?

1. Models of network structure
2. Power-laws and scale-free networks, “rich-get-richer” phenomena
3. Triadic closure and “the strength of weak ties”
4. Small-world phenomena
5. Hubs & Authorities; PageRank
CSE 158 – Lecture 12
Web Mining and Recommender Systems

Small-world phenomena
Small worlds

• We’ve seen random graph models that reproduce the **power-law** behaviour of real-world networks

• But what about other types of network behaviour, e.g. can we develop a random graph model that reproduces small-world phenomena? Or which have the correct ratio of closed to open triangles?
Small worlds

Social networks are **small worlds**: (almost) any node can reach any other node by following only a few hops

(picture from readingeagle.com)
Another famous study...

- Stanley Milgram wanted to test the (already popular) hypothesis that people in social networks are separated by only a small number of “hops”
- He conducted the following experiment:

1. “Random” pairs of users were chosen, with start points in Omaha & Wichita, and endpoints in Boston
2. Users at the start point were sent a letter describing the study: they were to get the letter to the endpoint, but only by contacting somebody with whom they had a direct connection
3. So, either they sent the letter directly, or they wrote their name on it and passed it on to somebody they believed had a high likelihood of knowing the target (they also mailed the researchers so that they could track the progress of the letters)
Another famous study...

Of those letters that reached their destination, the average path length was between 5.5 and 6 (thus the origin of the expression). At least two facts about this study are somewhat remarkable:

• First, that short paths appear to be abundant in the network
• Second, that people are capable of discovering them in a “decentralized” fashion, i.e., they’re somehow good at “guessing” which links will be closer to the target
Six degrees of separation

Such small-world phenomena turn out to be abundant in a variety of network settings

e.g. Erdos numbers:

<table>
<thead>
<tr>
<th>Erdös #</th>
<th>People</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 person</td>
</tr>
<tr>
<td>1</td>
<td>504 people</td>
</tr>
<tr>
<td>2</td>
<td>6593 people</td>
</tr>
<tr>
<td>3</td>
<td>33605 people</td>
</tr>
<tr>
<td>4</td>
<td>83642 people</td>
</tr>
<tr>
<td>5</td>
<td>87760 people</td>
</tr>
<tr>
<td>6</td>
<td>40014 people</td>
</tr>
<tr>
<td>7</td>
<td>11591 people</td>
</tr>
<tr>
<td>8</td>
<td>3146 people</td>
</tr>
<tr>
<td>9</td>
<td>819 people</td>
</tr>
<tr>
<td>10</td>
<td>244 people</td>
</tr>
<tr>
<td>11</td>
<td>68 people</td>
</tr>
<tr>
<td>12</td>
<td>23 people</td>
</tr>
<tr>
<td>13</td>
<td>5 people</td>
</tr>
</tbody>
</table>

http://www.oakland.edu/enp/trivia/
Six degrees of separation

Such small-world phenomena turn out to be abundant in a variety of network settings

e.g. Bacon numbers:
Six degrees of separation

Such small-world phenomena turn out to be abundant in a variety of network settings

Bacon/Erdos numbers:

Kevin Bacon → Sarah Michelle Gellar → Natalie Portman → Abigail Baird → Michael Gazzaniga → J. Victor → Joseph Gillis → Paul Erdos
Six degrees of separation

Dodds, Muhamed, & Watts repeated Milgram’s experiments using e-mail

- 18 “targets” in 13 countries
- 60,000+ participants across 24,133 chains
- Only 384 (!) reached their targets

Histogram of (completed) chain lengths – average is just 4.01!

<table>
<thead>
<tr>
<th>L</th>
<th>N</th>
<th>Location</th>
<th>Travel</th>
<th>Family</th>
<th>Work</th>
<th>Education</th>
<th>Friends</th>
<th>Cooperative</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19,718</td>
<td>33</td>
<td>16</td>
<td>11</td>
<td>16</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>7,414</td>
<td>40</td>
<td>11</td>
<td>11</td>
<td>19</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2,834</td>
<td>37</td>
<td>8</td>
<td>10</td>
<td>26</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1,014</td>
<td>33</td>
<td>6</td>
<td>7</td>
<td>31</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>349</td>
<td>27</td>
<td>3</td>
<td>6</td>
<td>38</td>
<td>12</td>
<td>6</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>117</td>
<td>21</td>
<td>3</td>
<td>5</td>
<td>42</td>
<td>15</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>37</td>
<td>16</td>
<td>3</td>
<td>3</td>
<td>46</td>
<td>19</td>
<td>8</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

**Reasons** for choosing the next recipient at each point in the chain

Six degrees of separation

Actual shortest-path distances are similar to those in Dodds’ experiment:

This suggests that people choose a reasonably good heuristic when choosing shortest paths in a decentralized fashion (assuming that FB is a good proxy for “real” social networks)

from “the anatomy of facebook”: http://goo.gl/H0bkWY
Q: is this result surprising?

• **Maybe not:** We have ~100 friends on Facebook, so $100^2$ friends-of-friends, $10^6$ at length three, $10^8$ at length four, **everyone** at length 5

• **But:** Due to our previous argument that people close triads, the vast majority of new links will be between friends of friends (i.e., we’re increasing the density of our local network, rather than making distant links more reachable)

• In fact 92% of new connections on Facebook are to a friend of a friend (Backstrom & Leskovec, 2011)
Six degrees of separation

**Definition:** Network diameter

- A network’s diameter is the length of its longest shortest path.
- **Note:** iterating over all pairs of nodes $i$ and $j$ and then running a shortest-paths algorithm is going to be prohibitively slow.
- Instead, the “all pairs shortest paths” algorithm computes all shortest paths simultaneously, and is more efficient ($O(N^2 \log N)$ to $O(N^3)$, depending on the graph structure).
- In practice, one doesn’t really care about the diameter, but rather the distribution of shortest path lengths, e.g., what is the average/90th percentile shortest-path distance.
- This latter quantity can computed just by randomly sampling pairs of nodes and computing their distance.
- When we say that a network exhibits the “small world phenomenon”, we are really saying this latter quantity is small.
Q: is this a contradiction?

- How can we have a network made up of dense communities that is simultaneously a small world?
- The shortest paths we could possibly have are $O(\log n)$ (assuming nodes have constant degree)

![random connectivity - low diameter, low clustering coefficient](http://cs224w.Stanford.edu)

![regular lattice - high clustering coefficient, high diameter](http://cs224w.Stanford.edu)
Six degrees of separation

We’d like a model that reproduces small-world phenomena

regular lattice – high clustering coefficient, high diameter

random connectivity – low diameter, low clustering coefficient

We’d like something “in between” that exhibits both of the desired properties (high cc, low diameter)

from [http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html](http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html)
Six degrees of separation

The following model was proposed by Watts & Strogatz (1998)

1. Start with a regular lattice graph (which we know to have high clustering coefficient)
   Next – introduce some randomness into the graph
2. For each edge, with prob. \( p \), reconnect one of its endpoints

as we increase \( p \), this becomes more like a random graph

from [http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html](http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html)
Six degrees of separation

Slightly simpler (to reason about formulation) with the same properties

1. Start with a regular lattice graph (which we know to have high clustering coefficient)
2. From each node, add an additional random link
Six degrees of separation

Slightly simpler (to reason about formulation) with the same properties

Conceptually, if we combine groups of adjacent nodes into “supernodes”, then what we have formed is a **4-regular** random graph

(very handwavy) proof:
- The clustering coefficient is still high (each node is incident to 12 triangles)
- **4-regular** random graphs have diameter $O(\log n)$ (Bollobas, 2001), so the whole graph has diameter $O(\log n)$

connections between supernodes:

(should be a 4-regular random graph, I didn’t finish drawing the edges)
Six degrees of separation

The Watts-Strogatz model

• Helps us to understand the relationship between dense clustering and the small-world phenomenon
• Reproduces the small-world structure of realistic networks
• Does **not** lead to the correct degree distribution (no power laws)
Six degrees of separation

So far...

• Real networks exhibit **small-world** phenomena: the average distance between nodes grows only logarithmically with the size of the network.
• Many experiments have demonstrated this to be true, in mail networks, e-mail networks, and on Facebook etc.
• But we know that social networks are highly **clustered** which is somehow inconsistent with the notion of having low diameter.
• To explain this apparent contradiction, we can model networks as some combination of highly-clustered nodes, plus some fraction of “random” connections.
Further reading:

- Easley & Kleinberg, Chapter 20
  - Milgram’s paper
    “An experimental study of the small world problem”
- Dodds et al.’s small worlds paper
- Facebook’s small worlds paper
  http://arxiv.org/abs/1111.4503
- Watts & Strogatz small worlds model
  “Collective dynamics of ‘small world’ networks”
  file:///C:/Users/julian/Downloads/w_s_NATURE_0.pdf
- More about random graphs
  “Random Graphs” (Bollobas, 2001), Cambridge University Press
CSE 158 – Lecture 12
Web Mining and Recommender Systems

Hubs and Authorities; PageRank
Trust in networks

We already know that there’s considerable variation in the connectivity structure of nodes in networks.

So how can we find nodes that are in some sense “important” or “authoritative”?

• In links?
• Out links?
• Quality of content?
• Quality of linking pages?
• etc.
1. The “HITS” algorithm

Two important notions:

**Hubs:**
We might consider a node to be of “high quality” if it links to many high-quality nodes. E.g. a high-quality page might be a “hub” for good content (e.g. Wikipedia lists)

**Authorities:**
We might consider a node to be of high quality if many high-quality nodes link to it (e.g. the homepage of a popular newspaper)
This “self-reinforcing” notion is the idea behind the HITS algorithm

- Each node $i$ has a “hub” score $h_i$
- Each node $i$ has an “authority” score $a_i$

- The hub score of a page is the sum of the authority scores of pages it links to
- The authority score of a page is the sum of hub scores of pages that link to it
This “self-reinforcing” notion is the idea behind the HITS algorithm

Algorithm:

\[ a_i^{(0)} = \frac{1}{\sqrt{n}} \quad h_i^{(0)} = \frac{1}{\sqrt{n}} \]

iterate until convergence:

\[ \forall_i a_i^{(t+1)} = \sum_{j \rightarrow i} h_j^{(t)} \]

\[ \forall_i h_i^{(t+1)} = \sum_{i \rightarrow j} a_j^{(t)} \]

normalize:

\[ \|a^{(t+1)}\|_2^2 = 1 \quad \|h^{(t+1)}\|_2^2 = 1 \]
Trust in networks

This “self-reinforcing” notion is the idea behind the HITS algorithm

This can be re-written in terms of the adjacency matrix ($A$)

$$a_i^{(0)} = \frac{1}{\sqrt{n}} \quad h_i^{(0)} = \frac{1}{\sqrt{n}}$$

iterate until convergence:

$$a^{(t+1)} = A^T h^{(t)}$$

skipping a step:

$$a^{(t+2)} = (A^T A)^t a^{(t)}$$

$$h^{(t+1)} = A a^{(t)}$$

$$h^{(t+2)} = (A A^T)^t h^{(t)}$$

normalize:

$$\| a^{(t+1)} \|_2^2 = 1 \quad \| h^{(t+1)} \|_2^2 = 1$$
Trust in networks

This “self-reinforcing” notion is the idea behind the HITS algorithm

So at convergence we seek stationary points such that

\[ A^T A a = c' \cdot a \]
\[ A A^T h = c'' \cdot h \]

(constants don’t matter since we’re normalizing)

- This can only be true if the authority/hub scores are eigenvectors of \( A^T A \) and \( A A^T \)
- In fact this will converge to the eigenvector with the largest eigenvalue (see: Perron-Frobenius theorem)
The idea behind PageRank is very similar:

- Every page gets to “vote” on other pages
- Each page’s votes are proportional to that page’s importance
- If a page of importance $x$ has $n$ outgoing links, then each of its votes is worth $x/n$
- Similar to the previous algorithm, but with only a single a term to be updated (the rank $r_i$ of a page $i$)

$$\forall_i r_i^{(t+1)} = \sum_j \frac{r_j^{(t)}}{|\Gamma(j)|}$$

(rank of linking pages)

(# of links from linking pages)
The idea behind PageRank is very similar:

Matrix formulation:
each \textbf{column} describes the out-links of one page, e.g.:

\[ M = \begin{pmatrix}
\frac{1}{3} & 0 & \frac{1}{4} & 1 \\
0 & 0 & \frac{1}{4} & 0 \\
\frac{1}{3} & \frac{1}{2} & \frac{1}{4} & 0 \\
\frac{1}{3} & \frac{1}{2} & \frac{1}{4} & 0 \\
\end{pmatrix} \]

column-stochastic matrix (columns add to 1)

this out-link gets 1/3 votes since this page has three out-links
The idea behind PageRank is very similar:

Then the update equations become:

$$r(t+1) = Mr(t)$$

And as before the stationary point is given by the eigenvector of $M$ with the highest eigenvalue.
Summary

The level of “authoritativenss” of a node in a network should somehow be defined in terms of the pages that link to (it or the pages it links from), and their level of authoritativenss

• Both the HITS algorithm and PageRank are based on this type of “self-reinforcing” notion
• We can then measure the centrality of nodes by some iterative update scheme which converges to a stationary point of this recursive definition
• In both cases, a solution was found by taking the principal eigenvector of some matrix encoding the link structure
This week

- We’ve seen how to characterize networks by their degree distribution (degree distributions in many real-world networks follow power laws)
- We’re seen some random graph models that try to mimic the degree distributions of real networks
- We’ve discussed the notion of “tie strength” in networks, and shown that edges are likely to form in “open” triads
- We’ve seen that real-world networks often have small diameter, and exhibit “small-world” phenomena
- We’ve seen (very quickly) two algorithms for measuring the “trustworthiness” or “authoritativeness” of nodes in networks
Further reading:

• Easley & Kleinberg, Chapter 14
• The “HITS” algorithm (aka “Hubs and Authorities”) “Hubs, authorities, and communities” (Kleinberg, 1999)

CSE 158 – Lecture 12
Web Mining and Recommender Systems

Some midterm Qs
Section 1: Regression and Ranking (6 marks)

Unless specified otherwise questions are each worth 1 mark.

1. The following is a list of prices from a local car dealership:

<table>
<thead>
<tr>
<th>No.</th>
<th>Model</th>
<th>Luxury?</th>
<th>Year</th>
<th>MPG</th>
<th>Horsepower</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Acura MDX</td>
<td>Yes</td>
<td>2017</td>
<td>20</td>
<td>290</td>
<td>$50,000</td>
</tr>
<tr>
<td>2</td>
<td>Honda Accord</td>
<td>No</td>
<td>2017</td>
<td>25</td>
<td>190</td>
<td>$25,000</td>
</tr>
<tr>
<td>3</td>
<td>Honda Civic</td>
<td>No</td>
<td>2012</td>
<td>23</td>
<td>160</td>
<td>$10,000</td>
</tr>
<tr>
<td>4</td>
<td>Honda Civic</td>
<td>No</td>
<td>2016</td>
<td>24</td>
<td>170</td>
<td>$18,000</td>
</tr>
<tr>
<td>5</td>
<td>Nissan Altima</td>
<td>No</td>
<td>2016</td>
<td>30</td>
<td>180</td>
<td>$25,000</td>
</tr>
<tr>
<td>6</td>
<td>Acura MDX</td>
<td>Yes</td>
<td>2015</td>
<td>18</td>
<td>280</td>
<td>$38,000</td>
</tr>
<tr>
<td>7</td>
<td>Lexus RX350</td>
<td>Yes</td>
<td>2015</td>
<td>21</td>
<td>270</td>
<td>$40,000</td>
</tr>
<tr>
<td>8</td>
<td>Toyota Prius</td>
<td>No</td>
<td>2014</td>
<td>45</td>
<td>120</td>
<td>$28,000</td>
</tr>
<tr>
<td>9</td>
<td>Toyota Prius</td>
<td>No</td>
<td>2013</td>
<td>40</td>
<td>120</td>
<td>$24,000</td>
</tr>
</tbody>
</table>

Suppose you train a regressor of the following form to predict a vehicle’s price:

\[
\text{price} \approx \theta_0 + \theta_1[\text{Year}] + \theta_2[\text{MPG}] + \theta_3[\text{Is luxury?}]
\]

What would be the feature representation of the first two vehicles?

1:

\[1 \ 2017 \ 20 \ 1\]

2:

\[1 \ 2017 \ 25 \ 0\]
2. List two additional features that might be useful for predicting the price of a car, and how you would encode them (2 marks):

1: Is used? (Yes, No)

2: Country of manufacture (One hot)

3. Suppose that you train two predictors on similar data to predict the price and obtain:

\[
\text{Price}^{\text{(Predictor 1)}} = 40000 - 100 \times [\text{MPG}] \quad \text{Price}^{\text{(Predictor 2)}} = 30000 + 10000 \times [\text{Is luxury?}] + 100 \times [\text{MPG}]
\]

The coefficient for MPG is negative for the first predictor, but positive for the second. Can you provide a brief explanation / interpretation of why this could be the case?

A: Luxury cars have bad MPG, but are worth more, so without isolating luxury feature, it looks like MPG is bad
4. **(Hard)** In class we stated that the best possible constant predictor (i.e., \( y_i \approx \alpha \)) was to set \( \alpha \) to be the mean value of \( y \) (i.e., \( \alpha = \frac{1}{N} \sum_i y_i \)). Show that this is the case when minimizing the MSE (hint: compute the derivative of the MSE and find the critical point by solving for \( \alpha \)) (**2 marks**):

\[
\text{MSE} = \frac{1}{N} \sum_i (\alpha - y_i)^2 \quad \frac{\partial \text{MSE}}{\partial \alpha} = \frac{2}{N} \sum_i (\alpha - y_i)
\]

A:

\[
= 0 \quad \text{if} \quad \alpha = \frac{1}{N} \sum_i y_i
\]
Section 2: Classification and Diagnostics (8 marks)

Suppose you train two (linear) SVM classifiers, \( A \) and \( B \), which produce the following separation boundaries:

5. What is the performance of the two classifiers in terms of the following (you may leave your expressions unsimplified) (4 marks):

<table>
<thead>
<tr>
<th>Metric</th>
<th>Classifier A</th>
<th>Classifier B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td># True positives</td>
<td></td>
<td></td>
</tr>
<tr>
<td># True negatives</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6. Suppose you were using your classifier to rank e-mails from ‘important’ (positive label) to ‘not important.’ Which of the two classifiers would you prefer and why?

A: \[ \text{Precision} \leq 5 \]

B: \[ \text{Precision} \geq 4 \]

7. Imagine that the goal of a classifier is to predict whether a person is \( \geq 20 \) years old. Two features that might be predictive include (a) height, and (b) vocabulary size. Would a Naïve Bayes classifier be suitable to train a predictor based on these two features? Explain why or why not.

A: Yes. If I know age, then height tells me nothing about vocabulary size.
8. (Critical thinking) A trivial classifier that we did not cover in class is a nearest neighbor classifier. This classifier has no parameters, and simply classifies points in the test set based on their similarity to points in the training set. That is, given a point \( X_i \) that we wish to classify, we consider all \( X_j \) in the training set, and select the label of the nearest one:

\[
y_i = y_{\text{argmin}_j} \| X_i - X_j \|_2^2
\]

Describe two settings (e.g. applications, properties of datasets, computational resources available, etc.) in which the nearest neighbor classifier would be (1) preferable to logistic regression, and (2) less preferable than logistic regression (2 marks)

(1) Pts are close but hard to separate

A:

(2) Features have very different scales

gender = [height, income in $]
9. Consider running the clique percolation algorithm with $K = 3$ on the following graph (see pseudocode on final page of exam):

what are the communities found by the algorithm? (you can draw your solution directly on the graph)
10. Using the boxes below, draw examples of sets of 2-d point sets for which
   (a) PCA would be more appropriate than hierarchical clustering
   (b) Hierarchical clustering would be more appropriate than PCA
   (c) Neither hierarchical clustering nor PCA would be appropriate

   (3 marks)

11. For the examples above, describe a real pair of features that might be described by the points you drew.
   ((b) is provided as an example) (2 marks):

   (a) dimension 1: \text{Ask}
       dimension 2: \text{Smell}
   (b) dimension 1: \text{Latitude}
       dimension 2: \text{Longitude}
   (c) dimension 1: \text{Random}
Section 4: Recommender Systems (7 marks)

On a popular music streaming website, a few users have listened to the following music:

<table>
<thead>
<tr>
<th>Album</th>
<th>Listened?</th>
<th>Liked?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nathan</td>
<td>Thomas</td>
</tr>
<tr>
<td><em>Lana Del Ray</em></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><em>Born to Die</em></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><em>Ultraviolence</em></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Honeymoon</em></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>Lust for Life</em></td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

12. Suppose you want to determine which users are similar to each other in terms of their *listening* behavior. What would be an appropriate metric for determining users’ similarity, and which two users would be most similar under this metric (list multiple in case of a tie)? (2 marks)

A: *Jaccard*

N & IC : 3/4
13. Suppose you want to determine which users are similar to each other in terms of their preferences. What would be an appropriate metric for determining users' similarity, and which two users would be most similar under this metric (list multiple in case of a tie)? Describe how you handle the '?' entries (2 marks).

A: \[
\cos \theta \geq 0 \\
NAD: \frac{1}{\sqrt{5} \sqrt{3}} \quad NP: \frac{1}{\sqrt{5} \cdot 1}
\]

14. (Critical Thinking) Suppose you wanted to design a recommender system to suggest points of interest in a city based on users' past activities/behavior/etc. Describe what data you would collect from users, how you would model the problem, and any issues that make this problem different from those we saw in class (3 marks).

A: 
\[
data: (user, poi, timestamp) \\
poi: (lon, lat) \\
user: (age, budget, p(n visits poi \mid t, p(t=0))
\]