COMS 6998: Social Networks — Due on Wednesday, November 19th, 2014

Assignment #3 – Important Nodes and Groups

Xiao Zhu, Di Li (TAs) — A. Chaintreau (instructor)

Why there is three parts in this assignment: Each part fulfills one of the objectives of the class:

- **Manipulate concepts:** Getting Familiar with the technical concepts used in class, by reproducing similar arguments. Being proficient by manipulating the object to answer some small-size problem.

  You are expected to answer this question rigorously, the answer can be quite short as long as it contains all the required argument to justify your answer.
- **Experience the concepts in real case:** Being able to reproduce these concepts on real or synthetic data. Study their properties in real examples.
- **Connect the concepts to real-life:** Interpret a problem you find in light of the notions you have learned. Develop some critical eye to determine how the concepts introduced are useful in practice.

How to read this assignment : Exercise levels are indicated as follows

($\mapsto$) "elementary": the answer is not strictly speaking obvious, but it fits in a single sentence, and it is an immediate application of results covered in the lectures.

Use them as a checkpoint: it is strongly advised to go back to your notes if the answer to one of these questions does not come to you in a few minutes.

($\curvearrowright$) "intermediary": The answer to this question is not an immediate translation of results covered in class, it can be deduced from them with a reasonable effort.

Use them as a practice: how far are you from the answer? Do you still feel uncomfortable with some of the notions? which part could you complete quickly?

($\not\mapsto$) "tortuous": this question either requires an advanced notion, a proof that is long or inventive, or it is still open.

Use them as an inspiration: can you answer any of them? does it bring you to another problem that you can answer or study further? It is recommended to work on this question only AFTER you are done with the rest!

Part A — Manipulating the Concepts

Exercise 1: Manipulating Importance metrics (6pt)

Imagine you work in a small joint venture whose website contains only 6 webpages about various projects that points to each other as seen in Figure 1. You would like to create a new project and, assuming that Google follows the HITs algorithm, have it placed first whenever the name of your company is searched. We will neglect here the links from webpages of other website (which are longer to obtain, and are more difficult to control). We will also assume that, due to the competitive nature of projects in your company, you cannot ask any project to point to you.

1. Compute the score as obtained by the Hubs and Authorities after two complete iterations in the original network.

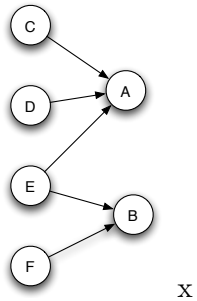

Figure 1: Elementary Network of Hubs and Authorities

2. Now imagine that you create a new project page X. Creating outgoing links from X is not going to help you getting a higher authority score. So you another fake project Y and points to X, as a way to provide some authority to X. You can also decide what project(s) Y links to.

   Compare the score you will obtain after two iterations in the HITs algorithm if Y points only to X or if Y points to A, B and X. Which one will you choose.

3. Imagine you can now create two fake projects Y and Z. Find a configuration that will place your project X as second in the order of importance. Again you will stop after two iterations.

Exercise 2: Joining an unbalanced world (6 pt)

In this exercise, we will learn whether a node can easily join an unbalanced world. We consider a graph $G = (V, E^+, E^-)$ that contains positive and negative edge and is complete (*i.e.*, all pairs of nodes have one edge between them). It is said *balanced* if all triangle contains either 1 or 3 friendship edges.

Imagine a new person X joins the graph G and has a complete freedom in deciding whether she maintains a positive or negative relationship with anybody else in the graph. The question is whether she can do so *while not being involved in any unbalanced triangle herself.* That means the following, if G is balanced, then X wants to join the graph while making sure she does not break this property. Otherwise, if G is already not balanced, some "forbidden triangles" exist, but at least X would like to make sure that after joining she is not part of any "forbidden triangles" herself.

1. ($\rightarrowtail$) Show by hand that if the graph G contains up to three nodes and is balanced, X can join the graph and not create any unbalanced triangle.

2. ($\curvearrowright$) Using a result from the class, prove that for a balanced graph of any size, then it is possible for X to always do so.

3. ($\curvearrowright$) Assuming now that G contains 3 nodes and is not balanced, what is the best thing that X can do to join the graph and not be part of any unbalanced triangle?

   For all cases of G, you should either prove that it is possible by providing a solution, or give a proof that it is impossible. We encourage that you provide a concise explanation instead of the result of an exhaustive search.

4. ($\rightarrowtail$) Provide a general conclusion describing in which cases X can successfully join any graph G and not be part of unbalanced triangle?

   Briefly discuss how this result generalizes when you use another definition of "balanced", generally called weakly balanced, in which triangles with three enemies are allowed, and the only triangles to avoid are the one with exactly two edges with positive signs.

Exercise 3: Game coordination and threshold model (0 pt, given for training)

Motivation: This exercice allows you to bridge simple coordination games, also referred to as cooperative game, with threshold model seen in the course. A coordination game is any type of game where players chooses from the same set of strategies and turn out to have higher payoff when they make the same choice.

Assume that each player can choose either strategy A or B. Strategy A may for example denote the adoption of an innovation. We suppose that players are connected to each other according to a graph $G = (V, E)$ where V denotes the set of all players and an edge (u, v) is in E if and only if two players are connected.

Every edge (u, v) corresponds to a game where players u and v are involved. In particular they all receive a pay-off for this game which only depends on which strategy each of this player has chosen according to the following table:

		v chooses A	v chooses B
u	chooses A	a	0
u	chooses B	0	b

where we assume that a and b are two real numbers such that $a > b$.

The total payoff receives by a user is the result of her payoff on all games she is involved in (*i.e.*, , she is involved by one game for each incident edge).

1. ($\rightarrowtail$) Given that all players except u has decided their strategy (which could either be A or B), and that u knows their choices, under which conditions will u rationally choose to play strategy A or B?

We consider an infinite sequence of games indexed by time $t = 0, 1, 2, \ldots$. We assume that at time t a user decides her strategy to maximize her payoff when all other players plays the same strategy as at time $t - 1$ (which is her last observation).

2. ($\rightarrowtail$) Let us assume that a set of players always choose to play strategy A independently of others, and that all other players initially play strategy B. How would you describe the evolution of this system with time t? Can you compare it to one seen in the course?

Exercise 4: Issues with Basic Pagerank (0.5pt)

Basic pagerank is the first version of pagerank, without a restart or, equivalently, with no "dumping factor". The importance metric of a node is hence the probability that a random walk in steady state is currently visiting this node. We have seen that this can create black hole whenever a node (or a small group of nodes) only receives links from other, but this does not seem to be the case if the graph is undirected.

1. ($\looparrowright$) Prove that basic pagerank on an undirected graph is not intrinsically better since it is equivalent to another metric we have seen.

2. (no credit, just for fun) Can you imagine a situation in which this algorithm is still useful?

Exercise 5: Adoption with neighbor effect and renewed decision (0 pt, given for training)

Motivation Adoption of an innovation (like an online service) could be promoted by encouraging users to start the service for free. One can distinguish a permanent promotion (where users could access the service for free forever) and temporary promotion where they have a free period. Clearly the first form of promotion (which is the one we studied in class) can only do better. On the other hand, and especially for service paid by subscription, the second option seems cheaper overall to organize. The exercise answers the following question "Could this form of promotion be significantly less efficient?"

In this exercise, we propose to show that, according to macroscopic metric (*i.e.*, , the ability for a finite set of players to create an infinite cascade of adoption), the two are equivalent.

As in the previous exercice, we consider a set of users V that are connected together along edges of a graph $G = (V, E)$. We consider an infinite amount of time slots $t = 0, 1, 2, \ldots$. We assume that all users have an adoption threshold θ which characterizes their behavior as follows: during a time slot t, a user observes how many of her friends used the service and renew her subscription for time slot $t + 1$ if only if at least a fraction θ of her friend have used the service during time slot t.

In the *temporary promotion model* we consider an extreme version where initially a set of users S_0 are proposed to use the service for free *for a single time slot*. At the end of this time slot, they may or may not renew the service depending on what they have observed and the threshold rule defined above, just like any other nodes in the network. The only effect of the promotion is to increase the set of users during the first time slot.

In the *targeted permanent promotion model* we assume that an initial set of users S_0 are proposed the service for free *for an unlimited amount of time*, while all other users who may use the service or not decides to do so according to the threshold rule defined above.

In the *general permanent promotion model*, we assume that any user who decides to use the service once will receive a free subscription in all subsequent time slot. All users who have never used the service may decide to adopt it or not according to the threshold rule defined above.

Let us denote by S_t for $t = 0, 1, \ldots$ the set of users that decide to subscribe to the service during time slot t in the temporary promotion model. Let us define for each subset $A \subseteq V$ the function f_θ as

$$f_\theta(A) = \{\ v \in V \ \mid \ \text{at least a fraction } q \geq \theta \text{ of neighbors of } v \text{ are in } A \ \}\ .$$

1. ($\rightarrowtail$)Show that for any $t \geq 0$, $S_t = f_\theta^{(t)}(S_0)$ where $f_\theta^{(t)}$ denotes the function f_θ applied t times.

Let us denote respectively by S'_t and S''_t for $t = 0, 1, \ldots$ the set of users that decide to subscribe to the service during time slot t in the targeted permanent promotion model and the general permanent promotion model.

2. ($\curvearrowright$) Show that $S'_t = S''_t = g_\theta^{(t)}(S_0)$ where for any subset $A \subseteq V$, $g_\theta(A) = f_\theta(A) \cup A$. What can you deduce w.r.t. the efficiency of the general permanent promotion strategy?

We consider an infinite graph $G = (V, E)$ where all nodes have finite degree. For a given threshold θ, we say that a set S_0 is an infectious set for the temporary promotion model if, starting from S_0 the sequence S_t of nodes that subscribe the service eventually reach all nodes. Formally, S_0 is infectious if

$$\forall v \in V, \exists k \geq 0 \text{ such that } \forall t \geq k, v \in S_t \ .$$

Note that this definition is shown here for the temporary promotion model (*i.e.*, using the sequence $(S_t)_{t\geq 0}$) and that the same definition could be used with S'_t and S''_t to define infectious set in the two other models.

3. ($\curvearrowright$) Give an example of a graph $G = (V, E)$ and a set S_0 that is infectious for the general permanent promotion model but not for the temporary promotion model.

Let S_0 be a finite subset that is infectious for the general permanent promotion model. We define S^+ as the subset that contains all nodes in S_0 as well as all neighbors of nodes in S_0. Since S_0 is infectious for the general permanent promotion model, there exists t_0 such that $S^+ \subseteq S'_{t_0}$, where S'_t is the sequence of subscribing nodes starting from S_0.

4. ($\curvearrowright$) Show that the subset $T = S'_{t_0}$ is infectious for the temporary promotion model (*i.e.*, that the sequence $(S_t)_{t\geq 0}$ starting from T eventually contains the whole set).

Exercise 6: Connection between two general models of influence (0 pt, given for training)

Motivation In this exercice, we prove that the two general models of influence with random thresholds are, under a natural condition, equivalent.

As a quick reminder from the lecture, one can define more general model of influence in one of the two following manner:

- Define for any node $u \in V$ a function g_u taking value in $[0; 1]$ and which is defined on all subset of neighbors of u (*i.e.*, for any $S \subseteq N(U)$ we define a value $g_u(S) \in [0; 1]$).

  Node's behavior is then characterized as follows. First, we assume that a set S_0 of nodes initially adopt the service, and that for any node $v \in V$ there exists a threshold θ_v which chosen once for all in $[0; 1]$ according to a uniform distribution.

  Then, for any time slot t, if during this time slot t, the set of neighbors of v which adopt the service is S, v will adopt the service if and only if we have $\theta_v \leq g_v(S)$.

- In another model, we assume that for any u and v such that (u, v) is an edge in E there exists a function $p_v(u, .)$ which takes value in $[0; 1]$ and is defined on all subset of neighbors of v that do not contain u (*i.e.*, for any $S \subseteq N(v)$ such that $u \notin S$, we define a value $p_v(u, S)$).

  The behavior of the nodes is then described as follows. A subset S_0 of nodes initially adopt the service or the innovation at time $t = 0$.

  Then for any $t = 0, 1, \ldots,$ whenever a node u adopts the service for the first time during t, for any node v that is a neighbor of u it makes a single attempt during this time slot to influence v. Note that if v has already been influenced and use the service nothing will happen. Otherwise it indicates that all previous attempt to influence v has failed. We denote by S_v the set of all nodes who attempted to influence v before u. What happens then is the following. With probability $p_v(u, S_v)$ (chosen independently from the past) the attempt is successful and v starts using the service at time $t + 1$. Otherwise, and hence with a probability $(1 - p_v(u, S_v))$ the attempt does not succeed and hence u is added to S_v.

1. ($\rightarrowtail$) The second model (using function p_v) is called order independent if, for a node v, the probability that it adopts the service after an attempt by nodes $u_1, u_2, \ldots, u_k$ does not depend on the order of the sequence but only on the set $\{u_1, \ldots, u_k\}$.

   Write the probability that v adopts the service if a set S of nodes attempt to influence v. Provide an example where this probability is not order independent.

2. ($\curvearrowright$) We assume that function g_v are all monotone (*i.e.*, $g_v(S) \leq g_v(T)$ when $S \subseteq T$). Show that the dynamics of adoption defined by g is equivalent to that defined by p_v if p_v satisfies:

$$p_v(u, S) = \frac{g_v(S \cup \{u\}) - g_v(S)}{1 - g_v(S)} .$$

3. ($\curvearrowright$) Similarly, show that if the functions p_v for all v follow the order independent property, the dynamics defined by p_v is equivalent to one defined using g_v for a proper choice of g_v.

4. ($\not\rightarrowtail$) Note that g_v and $p_v(u, .)$ for any v and any u are real valued functions defined on set. Hence, they may or may not be submodular (according to the definition seen in the class), and it makes sense to say that the first model (resp. the second) is submodular if and only if all functions g_v (resp. all functions $p_v(u, .)$) are submodular.

   We have just seen that the two models are equivalent in the sense that for any model defined with $(g_v)_{v \in V}$ there is an equivalent dynamics that can be defined using $(p_v(u, .))_{u,v \in V}$. So it does not really matter which definitions is used.

   Does this imply that any dynamics that is submodular for the first model is also submodular for the second model? Provide either a proof or a counterexample for each inclusion.

Part B — Experiencing the Concepts

Exercise 7: Data Challenge 2 Read and complete the first data challenge (see document attached), different due date.

Part C — Concepts at large

Get ready for your final project