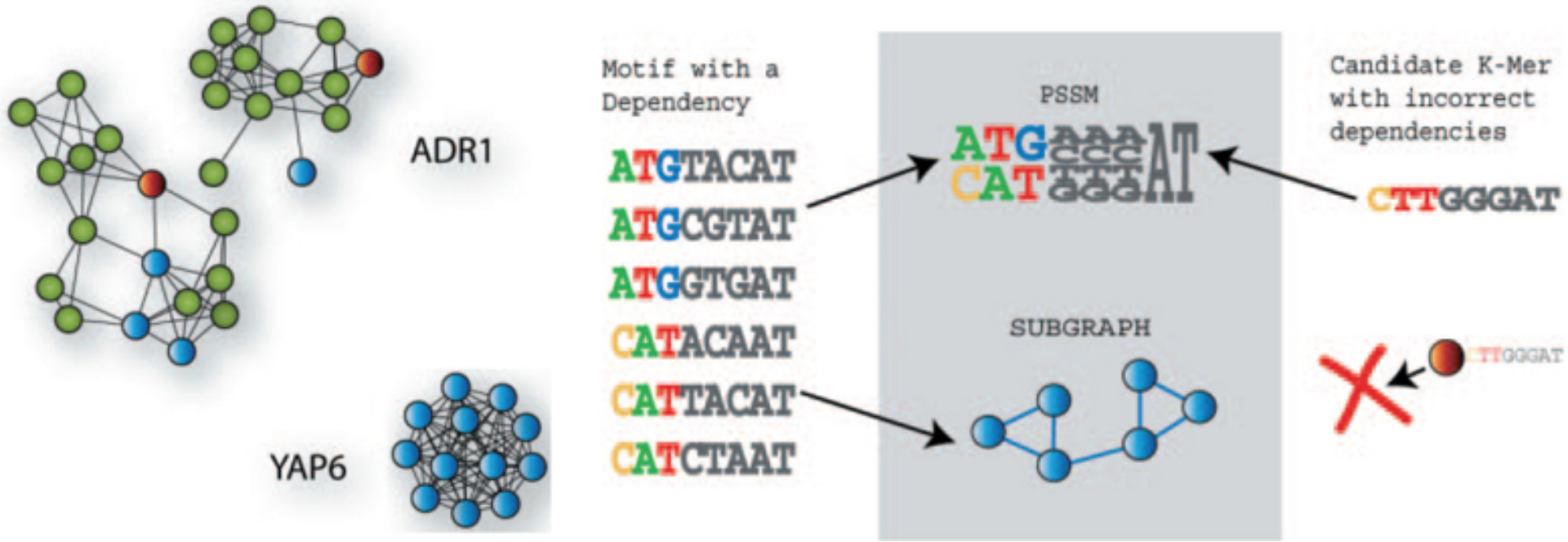
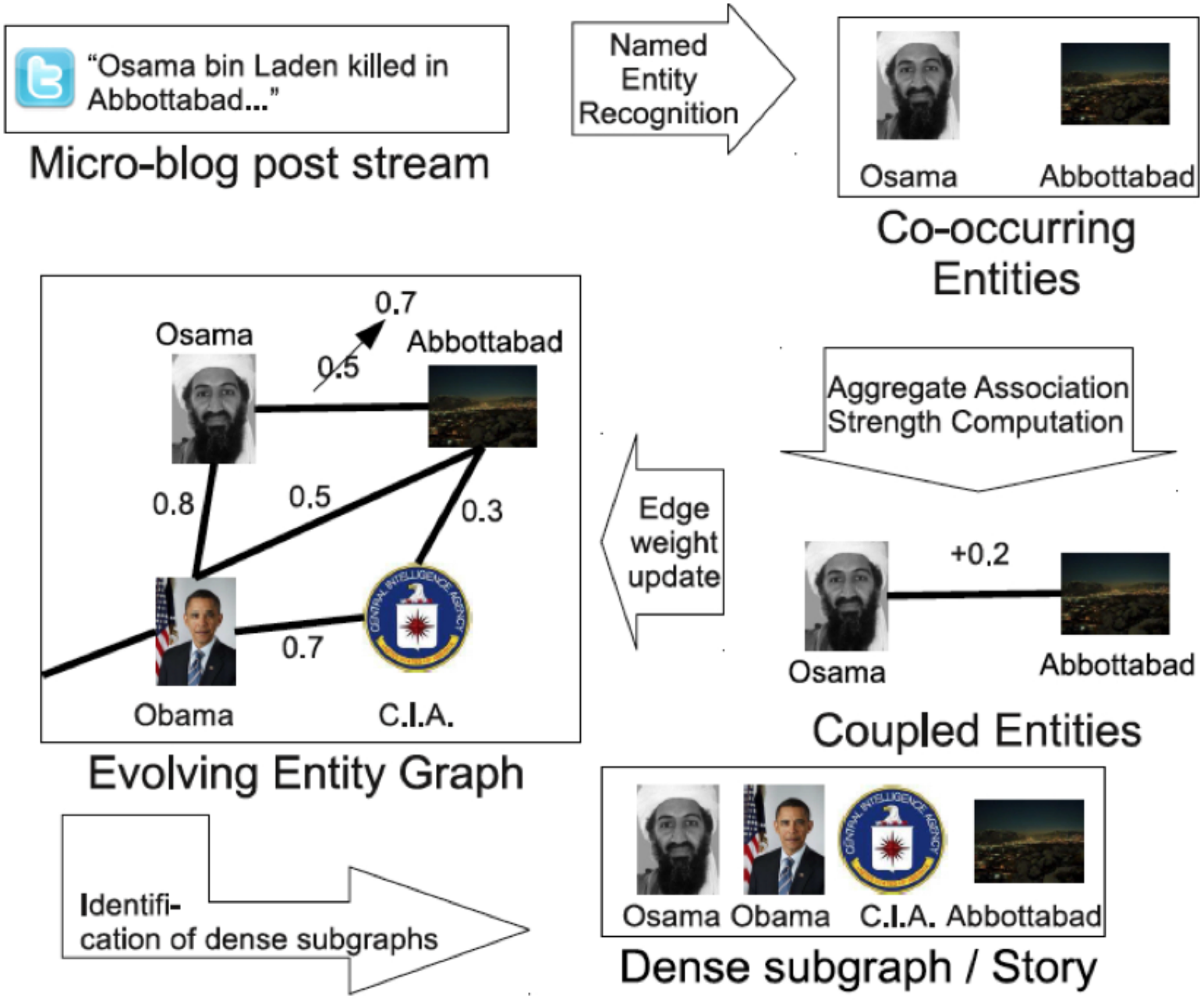
Influence Maximization Problem (IMP)

GRAPHS


ADR1
YAP6
Motif with a Dependency
ATGTACAT
ATGCGTAT
ATGGTGAT
CATACAAT
CATTACAT
CATCTAAT
PSSM
SUBGRAPH
Candidate K-Mer with incorrect dependencies
CTTGGGAT


"Osama bin Laden killed in Abbottabad..."
Micro-blog post stream
Named Entity Recognition
Osama
Abbottabad
Co-occurring Entities
0.7
Osama
Abbottabad
0.5
0.8
0.5
0.3
0.7
Obama
C.I.A.
Evolving Entity Graph
Aggregate Association Strength Computation
Edge weight update
+0.2
Osama
Abbottabad
Coupled Entities
Identifi-cation of dense subgraphs
Osama
Obama
C.I.A.
Abbottabad
Dense subgraph / Story

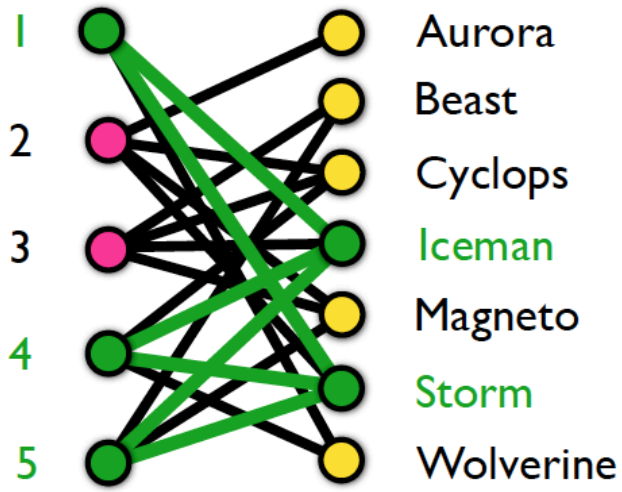
Graphs

id	heroes
1	Iceman, Storm, Wolverine
2	Aurora, Cyclops, Magneto, Storm
3	Beast, Cyclops, Iceman, Magneto
4	Cyclops, Iceman, Storm, Wolverine
5	Beast, Iceman, Magneto, Storm

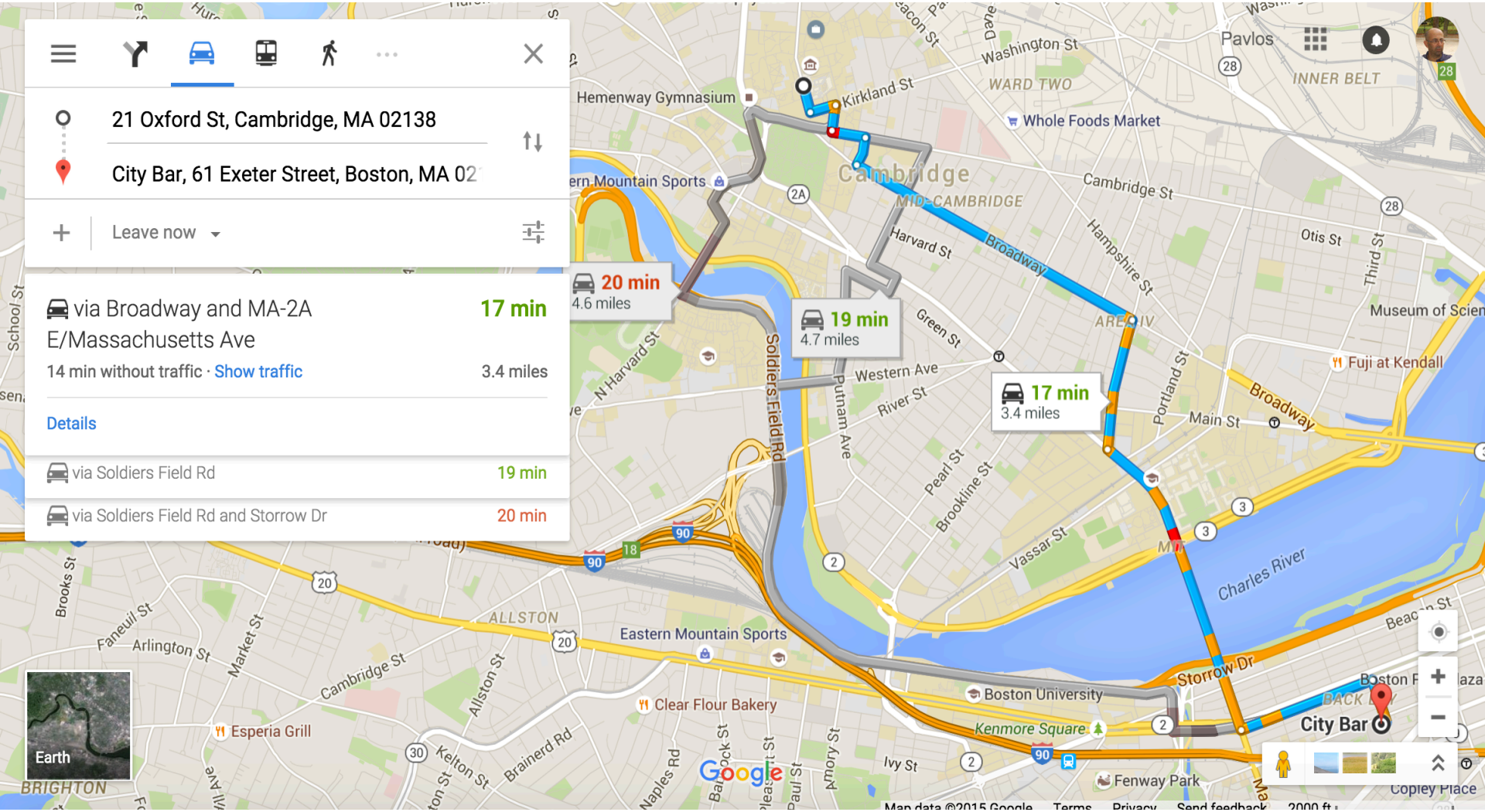
⇔

	A	B	C	I	M	S	W
1	0	0	0	1	0	1	1
2	1	0	1	1	1	0	0
3	0	1	1	1	1	0	0
4	0	0	1	1	0	1	1
5	0	1	0	1	1	1	0

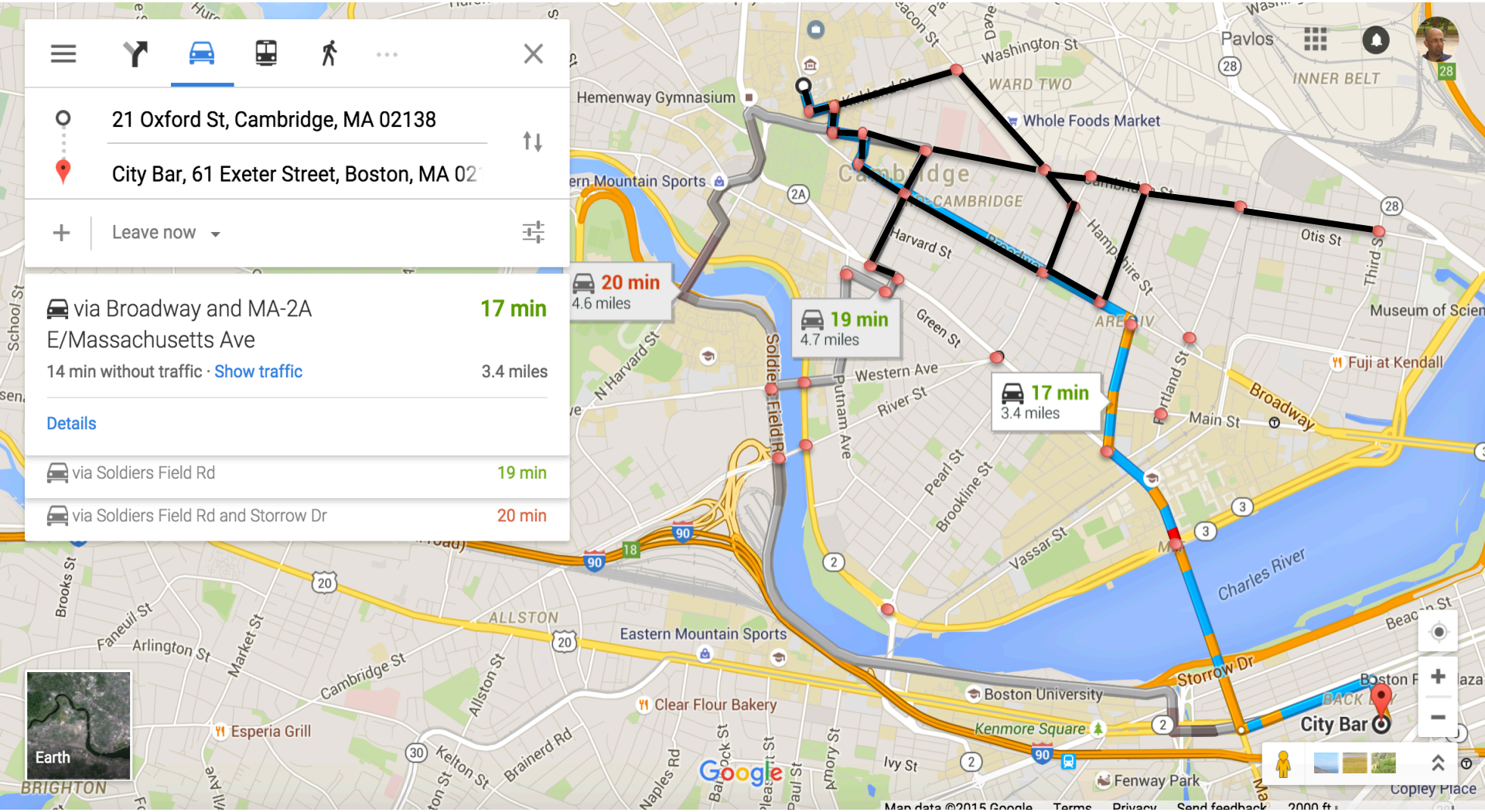
⇔



Driving to my favorite bar



More details



Social Network and Spread of Influence

- Social networks (traditional or digital) is THE medium for the spread of INFLUENCE among its members
  - Opinions, ideas, information, innovation…

- "Word-of-mouth" has been around since the Babylonians or the Greeks (Homer)
- Then came digital "Word-of-mouth" Gmail, Tupperware popularization, Facebook, Twitter)

The problem

- Given
  - a limited budget B for initial advertising (e.g. give away free samples of product)
  - influence between individuals (this can be probabilistic or deterministic)
- Goal
  - Create a large cascade of influence (e.g. more people know of the product)
- Question
  - Which set of individuals should B target at?
- Application besides product marketing
  - spread an innovation
  - stories in blogs

How we go about it

- Form models of influence in networks.
- Obtain data about particular network (to estimate inter-personal influence).
- Devise algorithm to maximize spread of influence.

Models of influence

- There are two main approaches
  - Linear Threshold
  - **Independent Cascade**
- First mathematical models in the 70s
  - Schelling 70/78, Granovetter 78
  - [Rogers 95, Valente 95, Wasserman 94
- A social network is represented:
  - as a directed graph with each person as a node
  - Nodes can be active or inactive
  - Active nodes may trigger neighbor nodes
  - [Active nodes never deactivate]

Linear Threshold

- A node *v* has random threshold $\vartheta_v \sim U[0,1]$
- A node *v* is influenced by each neighbor *w* according to a *weight* b_{vw} such that

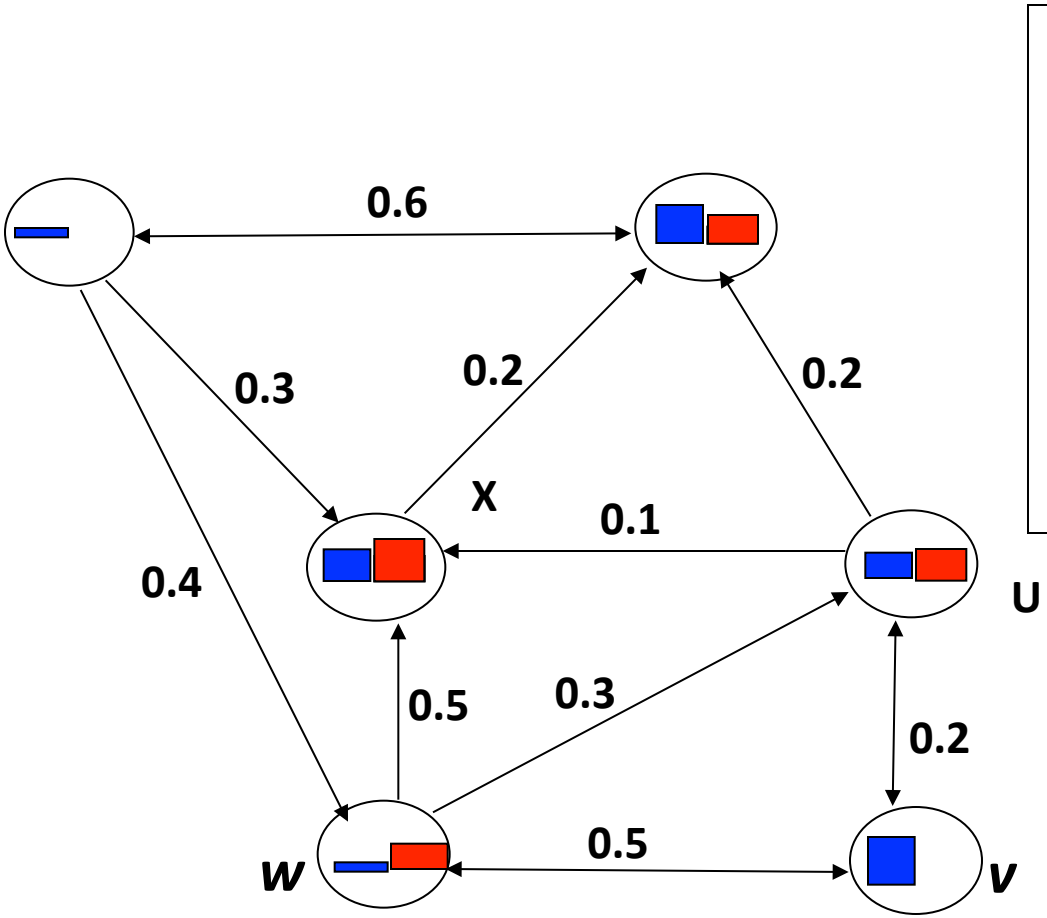
$$\sum_{w \text{ neighbor of } v} b_{v,w} \leq 1$$

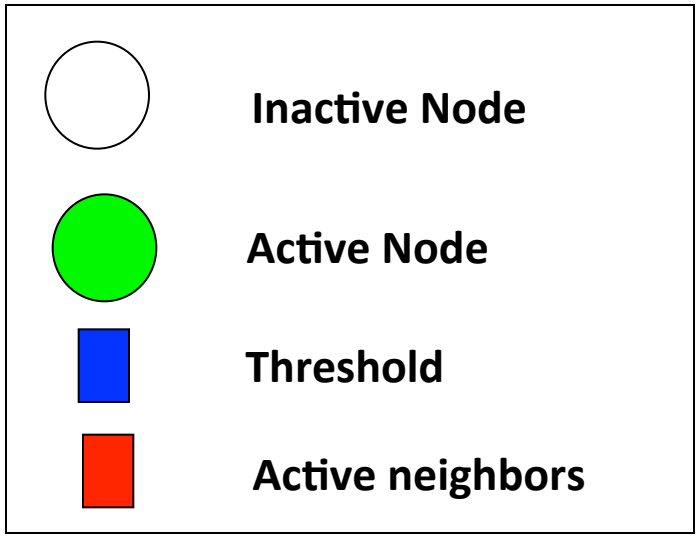
- A node *v* becomes active when at least

(weighted) ϑ_v fraction of its neighbors are active

$$\sum_{w \text{ active neighbor of } v} b_{v,w} \geq \theta_v$$

Example





Stop!

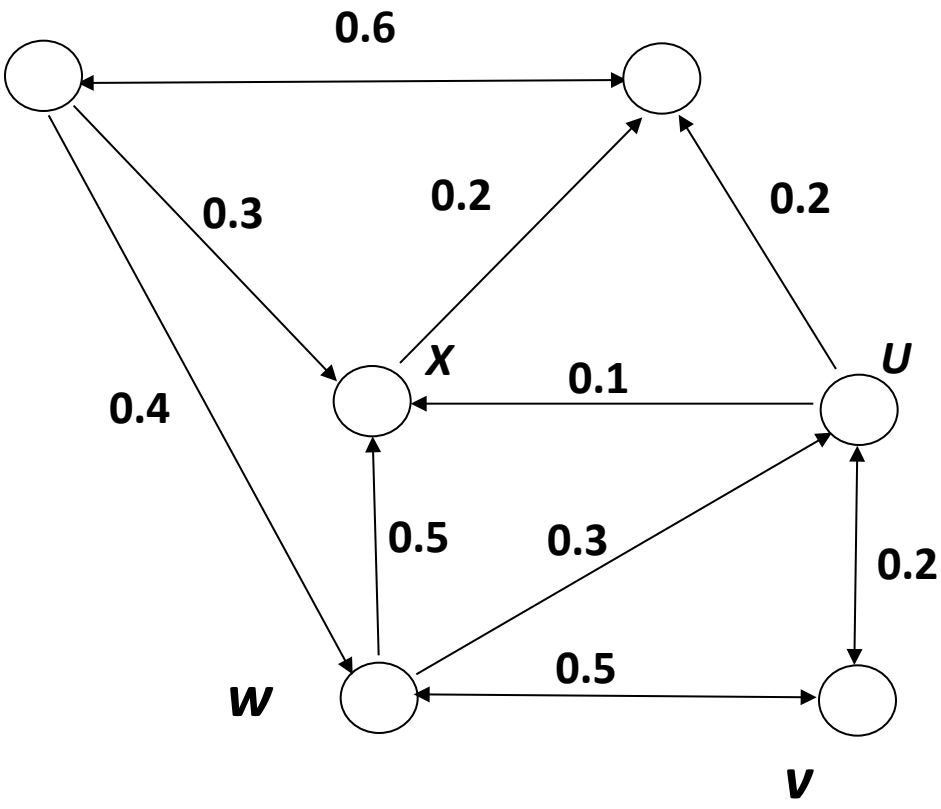
Independent Cascade

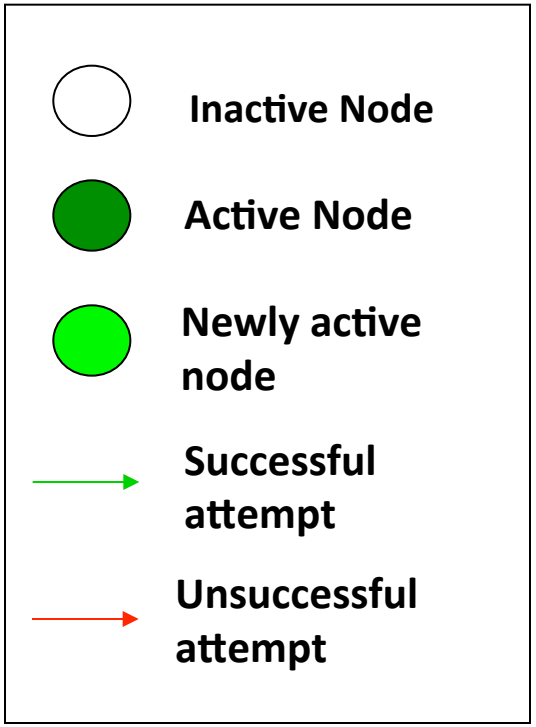
- An active node v has only one chance to activate inactive neighbors (currently inactive)
- The activation attempt succeeds with probability P_{vw}

Independent Cascade

- We again start with an initial set of active nodes A_0
- Process is in discrete steps
- When node v first becomes active in step t, it is given a single chance to activate each currently inactive neighbor w
  - With a probability Pvw —a parameter of the system — independently of the history thus far.
  - (If w has multiple newly activated neighbors, their attempts are sequenced in an arbitrary order.)
- If v succeeds, then w will become active in step t+1; but whether or not v succeeds, it cannot make any further attempts to activate w in subsequent rounds.
- The process runs until no more activations are possible.

Example





Stop!

Independent Cascade

- REMEMBER WE FLIP A COIN

Influence Maximization Problem

- Define the influence of node set S (this is a set of individuals): f(S)
  - Expected number of active nodes at the end given an initial set of active nodes S
- Problem:
  - Given an initial number of active nodes k, find a k-node set S to maximize f(S)
  - Optimization problem with f(S) as the objective function

Properties of f(S)

- Non-negative (dah)
- Monotone $f(S+v) \geq f(S)$

- Submodular:
  - Let N be a finite set
  - A set function f: 2^N

$$\forall S \subset T \subset N, \forall v \in N \setminus T,$$
$$f(S+v) - f(S) \geq f(T+v) - f(T)$$

Bad news

- For a submodular function *f*, if *f* only takes non-negative value, and is monotone, finding a *k*-element set *S* for which *f(S)* is maximized is an NP-hard optimization problem[GFN77, NWF78].
- It is NP-hard to determine the optimum for influence maximization for both independent cascade model and linear threshold model.

Good news

- We can use Greedy Algorithm or **Stochastic Methods**!
- **Greedy**:
  - Start with an empty set S
  - For k iterations:

    Add node v to S that maximizes $f(S+v) - f(S)$.
- How good (bad) it is?
  - Theorem: The greedy algorithm is a $(1 - 1/e)$ approximation.
  - The resulting set S activates at least $(1- 1/e) > 63\%$ of the number of nodes that any size-k set S could activate.

Evaluating *f(S)*

- How to evaluate *f*(*S*)?
- Still an open question of how to compute efficiently
- But: very good estimates by simulation
  - repeating the diffusion process often enough
  - Achieve *(1± ε)*-approximation to *f(S)*.
  - WHAT IS *ε?*

Data

- A review graph obtained from Yelp data set challenge
- We have the whole US (>6,000 businesses)
- Resulting graph: >350,000 nodes, 4,000,000 distinct edges

Experiment Settings

- Independent Cascade Model:
  - Edge from *v* to *ω* has probability (Beta(α,β)) of activating *ω*.
  - We learn α,β using existing reviews
  - β total number of reviews a reviewer w has written
  - α is he total number of reviews reviewer w has written after reviewer v has written (within time τ)

  Mean: $\alpha/(\alpha+\beta)$

  Variance: $(\alpha\beta)/[(\alpha+\beta)^2(\alpha+\beta+1)]$

- Simulate the process N times for each targeted set, re-edge outcomes pseudo-randomly from [0, 1] every time

- Use simulating annealing to solve the problem

Experimental setting

- τ: Time lag (1 month)
- Graphs will be given to you
  - Big
  - Small (toy; North Carolina)
- Script that generates the graph is available but you will not need it

Graph

- Graph is in Json format (like a dictionary).
- You will use NetworkX to load it => dictionary

Probabilistic Model

- Flip a coin and if value is

Goals for today

Networks Graphs etc

Independent Cascade Model

Influence Maximization

Probabilistic Models

- Form teams
- Ensure everyone had anaconda installed and NetworkX
- Installing Vagrant etc