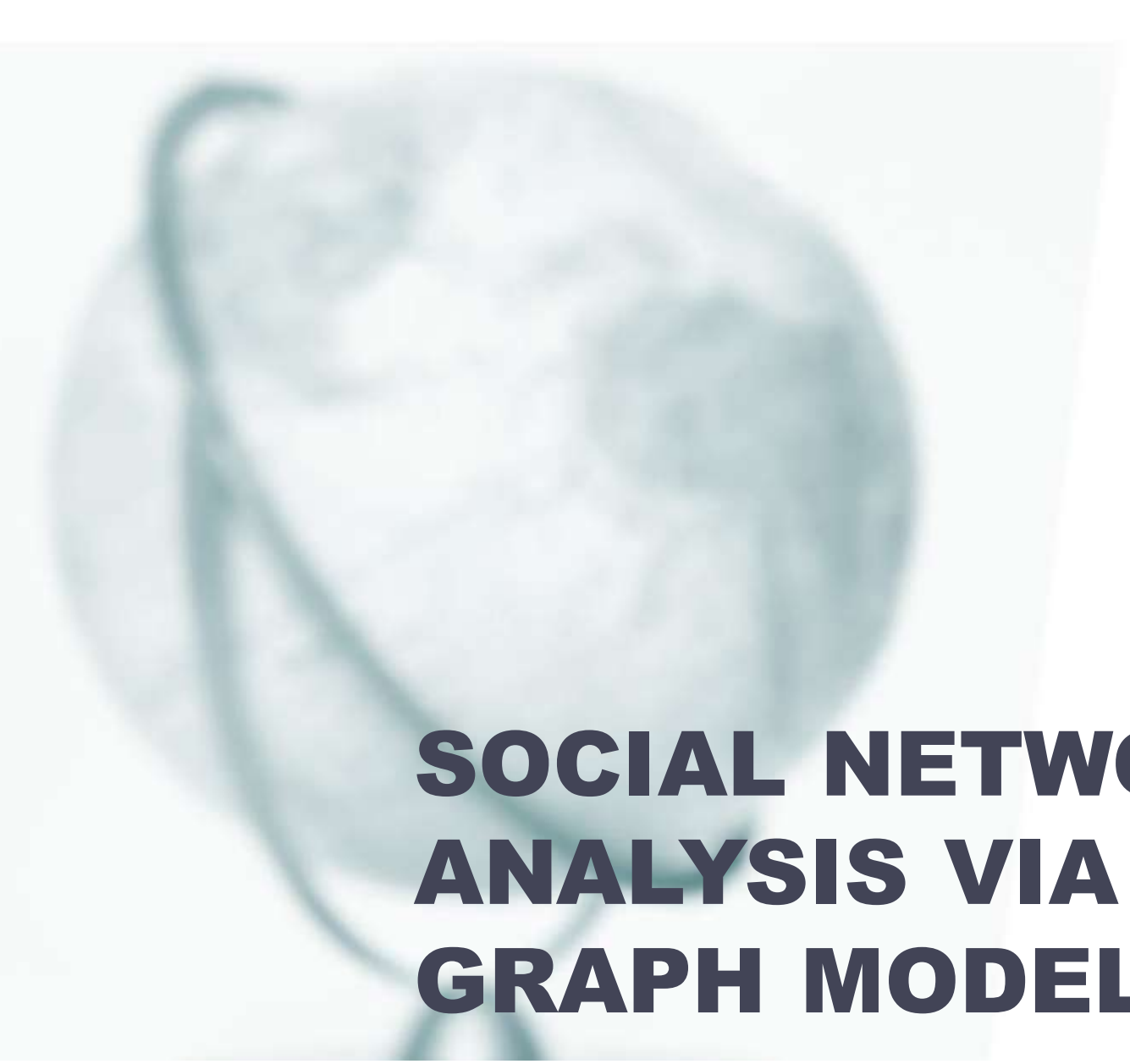


SOCIAL NETWORK ANALYSIS VIA FACTOR GRAPH MODEL

Zi Yang

OUTLINE

- Background
- Challenge
- Unsupervised case 1
  - Representative user finding
- Unsupervised case 2
  - Community discovery
- Experiments
- Supervised case
  - Modeling information diffusion in social network

BACKGROUND

- Social network

- Example: Digg.com
  - A popular social news website for people to discover and share content
  - Various types of behaviors of the users
    - submit, digg, comment and reply a comment
  - Edges
    - if one diggs or comments a story of another

BACKGROUND

- Community discovery
  - Modularity property

$$\exp \sum_{i,j} [y_i = y_j] \left(\alpha_{i,j} - \frac{k_i k_j}{2m} \right)$$

Pair-wise constrain

- Affinity propagation
  - Clustering via factor graph model
  - Update rules:

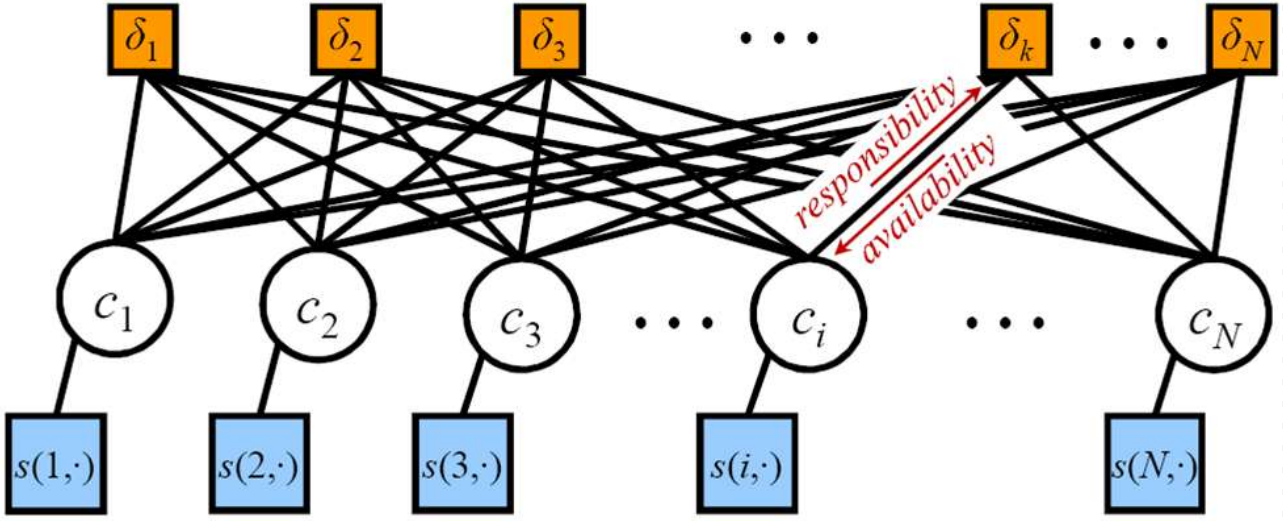
$$r(i,k) = s(i,k) - \max_{k's.t.k' \neq k} \{a(a,k') + s(i,k')\}$$

$$a(i,k) = \min\{0, r(k,k) + \sum_{i's.t.i' \notin \{i,k\}} \max\{0, r(i',k)\}\}$$

$$a(k,k) = \sum_{i's.t.i' \notin \{k\}} \max\{0, r(i',k)\}$$

BACKGROUND

- Affinity propagation



$$S(c) = \sum_{i=1}^{N} s(i, c_i) + \sum_{k=1}^{N} \delta_k(c_{1:N}) \quad \text{where} \quad \delta_k(c_{1:N}) = \begin{cases} -\infty, & \text{if } c_k \neq k \text{ but } \exists i : c_i = k \\ 0, & \text{otherwise} \end{cases}$$

Local factor

Regional constrain

OUTLINE

- Background
- Challenge
- Unsupervised case 1
  - Representative user finding
- Unsupervised case 2
  - Community discovery
- Experiments
- Supervised case
  - Modeling information diffusion in social network

CHALLENGES

- How to capture the local properties for social network analysis?
- Community discovery as a graph clustering, and how to consider the edge information directly?
  - Homophily
- What constraint can be applied to describe the formation/evolution of community?

OUTLINE

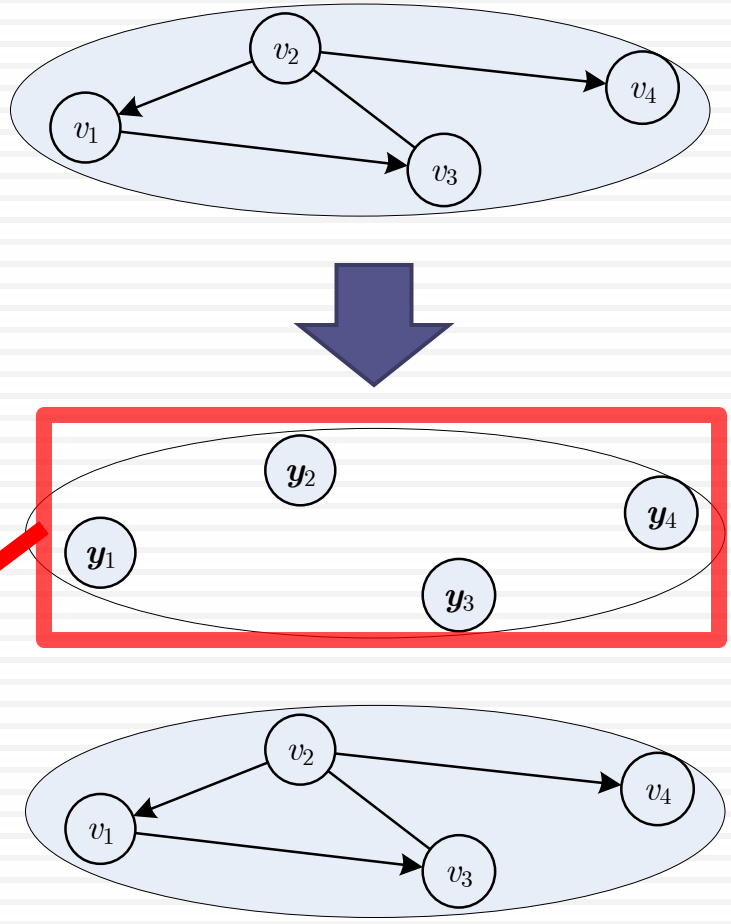
- Background
- Challenge
- Unsupervised case 1
  - Representative user finding
- Unsupervised case 2
  - Community discovery
- Experiments
- Supervised case
  - Modeling information diffusion in social network

REPRESENTATIVE USER FINDING

- Problem definition
  - given a social network $G = (V, E)$ and (optional) a confidence θ_i for each user v_i, the objective is to find a pair-wise representativeness on each edge in the network, and estimate the representative degree of each user v_i in the network, which is denoted by a set of variables $\{y_i\}$ satisfying $y_i \in \{1, \ldots, N\}$. In other words, y_i represents the user that v_i mostly trusts (or relies on).

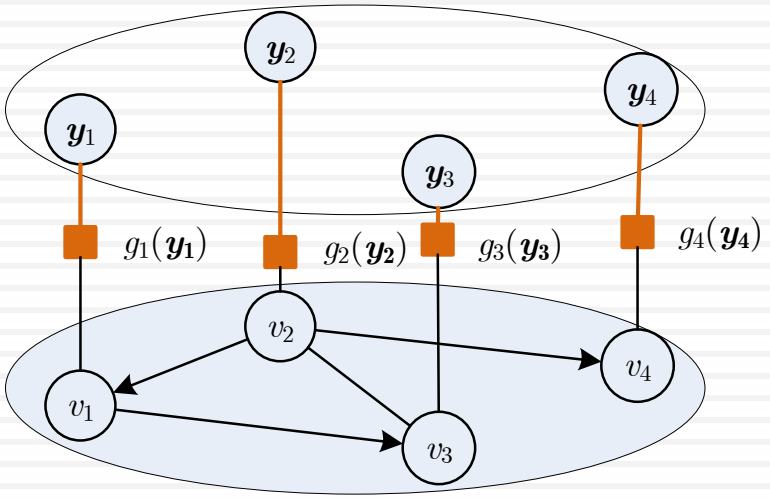
REPRESENTATIVE USER FINDING

- Modeling
  - Input
  - Variables



REPRESENTATIVE USER FINDING

- Modeling
  - Node feature function



Observation: similarity between the node and variable

Normalization factor

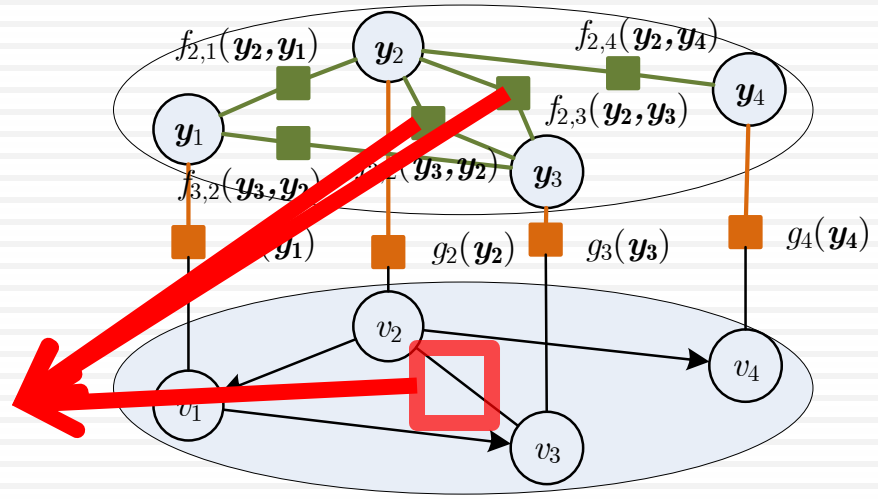
$$g_i(\mathbf{y_i}) = g_i(y_i) = \begin{cases} \kappa w_{i,y_i} & \text{if } y_i \in O(i) \\ \kappa \sum_{j \in NB(i)} w_{j,i} & \text{if } y_i = i \\ 0 & \text{otherwise} \end{cases}$$

Neighbor Representative

Self-representative

REPRESENTATIVE USER FINDING

- Modeling
  - Edge feature function



Undirected edge:
bidirected influence

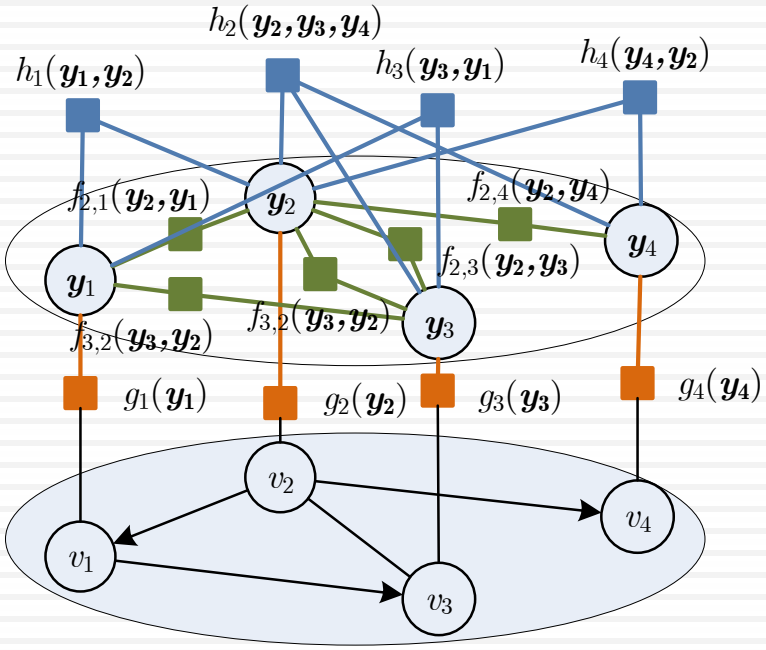
$$f_{i,j}(\mathbf{y}_i, \mathbf{y}_j) = f_{i,j}(y_i, y_j) = \begin{cases} \lambda & \text{if } y_i = y_j \\ 1-\lambda & \text{if } y_i \neq y_j \end{cases}$$

If vertexes of the edge have the same representative

If vertexes of the edge have different representative

REPRESENTATIVE USER FINDING

- Modeling
  - Regional feature function
    - a feature function defined on the set of neighboring nodes of v_i and itself.



$$h_k(\mathbf{y}_{\mathbf{I}(\mathbf{k})\cup\{\mathbf{k}\}}) = h_k(y_{I(k)\cup\{k\}}) = \begin{cases} 0 & \text{if } y_k = k \text{ and } \forall i \in I(k), y_i \neq k \\ 1 & \text{otherwise} \end{cases}$$

To avoid “leader without followers”

REPRESENTATIVE USER FINDING

- Modeling
  - Objective function

$$\max_{\mathbf{y}_{1:N}} \log P(\mathbf{y}_{1:N})$$

$$P(\mathbf{y}_{1:N}) = \frac{1}{Z}\prod_{i=1}^{N} g_i(\mathbf{y}_i)\prod_{e_{i,j}\in E} f_{i,j}(\mathbf{y}_i,\mathbf{y}_j)\prod_{k=1}^{N} h_k(\mathbf{y}_{\mathbf{I}(\mathbf{k})\cup\{\mathbf{k}\}})$$

$$= \frac{1}{Z}\left(\prod_{i=1}^{N} g_i(y_i)\prod_{e_{i,j}\in E} f_{i,j}(y_i,y_j)\prod_{k=1}^{N} h_k(y_{I(k)\cup\{k\}})\right)$$

- Solving
  - Max-sum algorithm

REPRESENTATIVE USER FINDING

- Model learning

$$a_{ii} = \max_{k \in I(j)} \min \left\{ r_{kj}, 0 \right\}$$

$$a_{ij} = \min \left\{ -\min \left\{ r_{jj}, 0 \right\} - \max_{k \in I(j) \setminus \{i\}} \min \left\{ r_{kj}, 0 \right\}, \max \left\{ r_{jj}, 0 \right\} \right\}$$

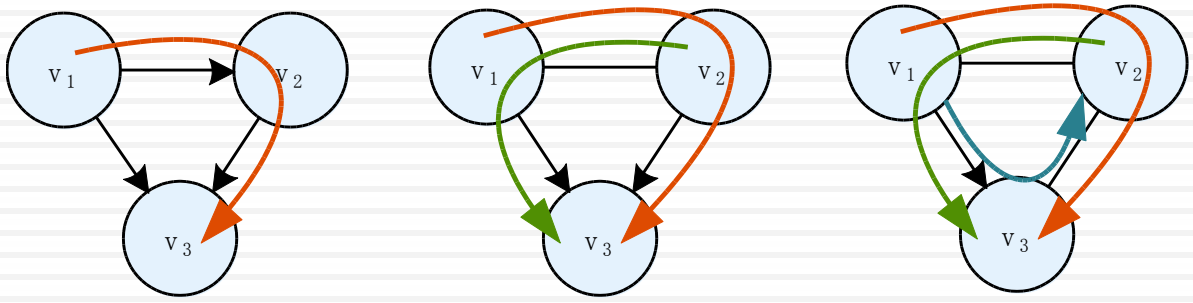
$$r_{ij} = g_{ij} + \sum_{k \in I(i) \cup O(i)} c_{ikj} - \max_{j' \in O(i) \cup \{i\} \setminus \{j\}} \left(g_{ij'} + a_{ij'} + \sum_{k \in I(i) \cup O(i)} c_{ikj'} \right)$$

$$p_{ijk} = g_{ik} + a_{ik} + \sum_{l \in I(i) \cup O(i) \setminus \{j\}} c_{ikl} - \max_{j' \in O(i)} \left(g_{ij'} + a_{ij'} + \sum_{l \in I(i) \cup O(i) \setminus \{j\}} c_{ilj'} \right)$$

$$c_{ijk} = \max \left\{ \log \frac{\lambda}{1-\lambda} + p_{jik}, 0 \right\}$$

REPRESENTATIVE USER FINDING

- A bit explanation
  - p_{ijk} : how likely user v_i persuades v_j to take v_k as his representative
  - c_{ijk} : how likely user v_i compliances the suggestion from v_j that he considers v_k as his representative
- The direction of such process
  - Along the directed edges



REPRESENTATIVE USER FINDING

- Algorithm

Algorithm 1: Representative user finding

input : A network $G = (V, E)$ and confidence $\{\Theta_v\}_{v \in V}$
output : Representativeness influence graph $G = (V, E)$

```
// initialize
foreach i and j ∈ O(i) ∪ {i} do
    Calculate g_ij according to Eq 10;
    Initialize r_ij ← −∞;
// update a_ij and r_ij on each node and edge
foreach i and j ∈ O(i) ∪ {i} do
    Update a_ij according to Eq 5 or Eq 6;
    Update r_ij according to Eq 7;
// update p_ijk and c_ijk on each edge and triangle
foreach i and j ∈ I(i) ∪ O(i) do
    foreach k ∈ (O(i) ∪ {i}) ∩ (O(j) ∪ {j}) do
        Update p_ijk according to Eq 8;
        Update c_ijk according to Eq 9;
```

OUTLINE

- Background
- Challenge
- Unsupervised case 1
  - Representative user finding
- Unsupervised case 2
  - Community discovery
- Experiments
- Supervised case
  - Modeling information diffusion in social network

COMMUNITY DISCOVERY

- Problem definition
  - given a social network G and an expected number of communities C , correspondingly a virtual node $u_c \in U$. is introduced for each community, and the objective is to find a community y_i for each person v_i satisfying $y_i \in \{1, \dots, C\}$, which represents the community that v_i belongs to, such that maximize the preservation of structure (or maximize the modularity Q of the community).

COMMUNITY DISCOVERY

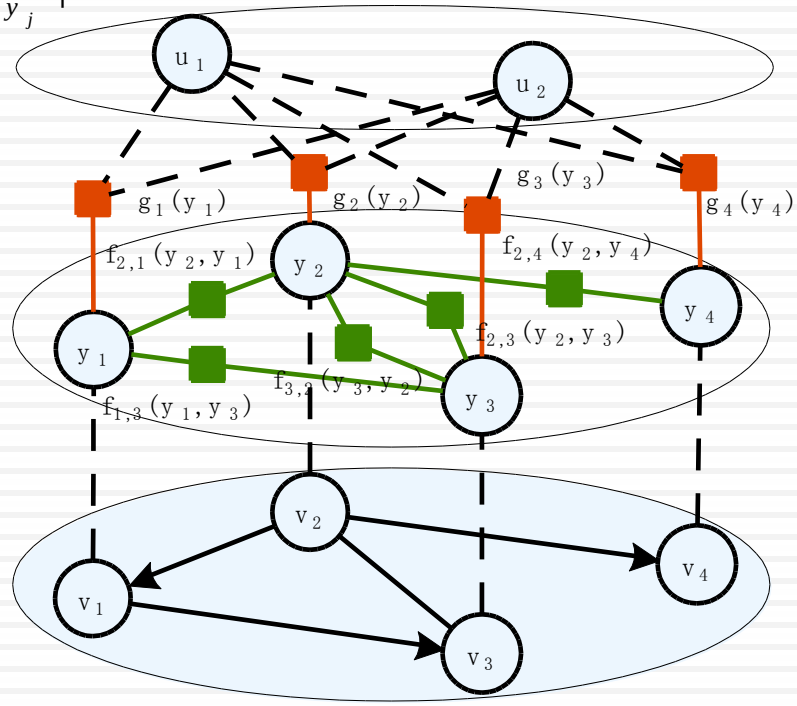
- Feature definition – What's different?
  - Node feature function

$$g_i(y_i) = \exp \sum_{j \in I(i) \cup O(i)} \left([y_j = y_i] + 1 \right) \frac{\alpha_{i,j}}{|X_{y_j}|}$$

  - Edge feature function

$$f_{i,j}(y_i, y_j) = \exp q_{i,j}$$

$$\propto \exp [y_i = y_j] \left(\alpha_{i,j} - \frac{k_i k_j}{2m} \right)$$



COMMUNITY DISCOVERY

- Algorithm

Algorithm 2: Algorithm for community discovery

input : A network $G = (V, E)$ and a specified number of communities C

output : Topic-level community influence graph $G = (V, E)$

```
// initialize
foreach i ← 1 to N do
    Randomly generate a community index c;
    y_i ← c, X_c ← X_c ∪ {i};
foreach i and j ∈ O(i) ∪ {i} do
    Calculate g_ij according to Eq 16;
    Initialize r_ij ← −∞;
// update p_ijk and c_ijk on each edge and triangle
foreach i and j ∈ I(i) ∪ O(i) do
    foreach k ∈ (O(i) ∪ {i}) ∩ (O(j) ∪ {j}) do
        Update p_ijk according to Eq 14;
        Update c_ijk according to Eq 15;
// update y_i and X_c for each node and edge
foreach i ← 1 to N do
    X_{y_i} ← X_{y_i} \ {i};
    c = argmax_c(p_ijk + c_ijk) for some j ∈ I(i) ∪ O(i);
    y_i ← c, X_c ← X_c ∪ {i};
```

Result output and Variable updates

OUTLINE

- Background
- Challenge
- Unsupervised case 1
  - Representative user finding
- Unsupervised case 2
  - Community discovery
- Experiments
- Supervised case
  - Modeling information diffusion in social network

Experiments

- Dataset: Digg.com
  - a popular social news website for people to discover and share content
  - 9,583 users, 56,440 contacts
  - various types of behaviors of the users
    - submit, digg, comment and reply a comment
  - Edges (In total: 308,362)
    - if one diggs or comments a story of another
    - Weight of the edge: the total number of diggs and comments

Experiments

- Dataset: Digg.com
  - 9,583 users, 56,440 contacts
  - 308,362 edges
    - weight of the edge: the total number of diggs and comments
- Settings:
  - Parameter $\alpha = 0.6$

Experiments

- Result: 3 most self-representative users on 3 different topics for Digg user network

DATASET	USER ID	TOPIC DISTRIBUTION			OUT DEGREE	IN DEGREE
		TOPIC 1	TOPIC 2	TOPIC 3		
Digg-1	puppyinlove	0.4444	0.0111	0.0111	5	0
	linkserf	0.2199	0.0015	0.0015	3	6
	chicagojack	0.0362	0.0012	0.0012	12	6
Digg-2	pyrates	0.0111	0.4444	0.0111	39	2
	GordonFree	0.0053	0.2128	0.0053	49	0
	maxthreepwood	0.0100	0.5000	0.0100	23	0
Digg-3	M724	0.0056	0.0056	0.2260	33	2
	Elliottx	0.0093	0.0093	0.0926	45	0
	3leggedHorse	0.0034	0.0034	0.1695	25	0

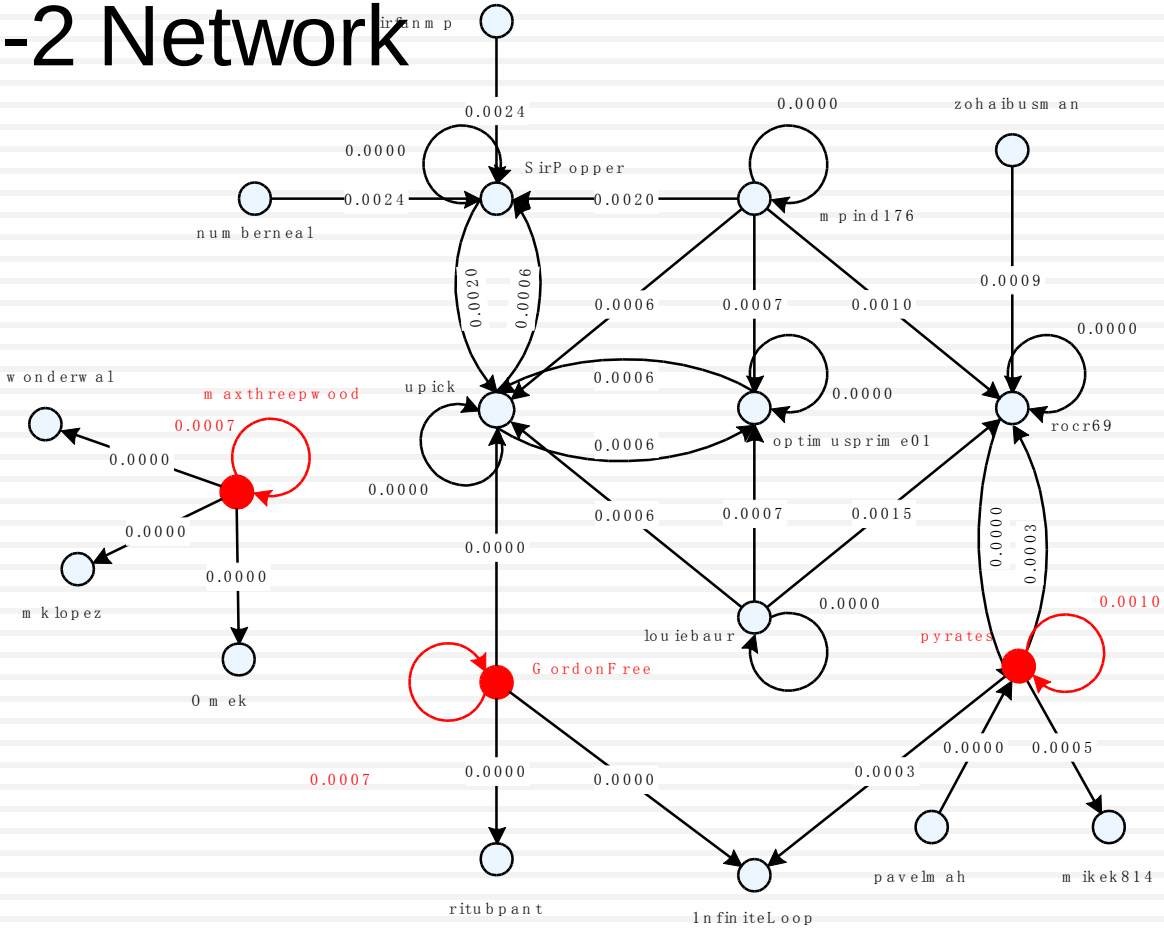
Experiments

- Result: 3 most representative users of 5 communities on 3 different subset

DATA SET	REPRESENTATIVE PERSONS
Digg-1	OuchLOL (0, 11), trentmonica (8, 4), chicagojack (12, 6)
	metejada (0, 7), nhuong (3, 4), JustinFallible (3, 7)
	llamarama7 (2, 1), kellfinder (19, 20), ethanator1088 (5, 2)
	AmyVernon (43, 46), diggleague (30, 43), DigSomeMore (39, 43)
	superman7018 (42, 31), dpratt356 (26, 31), Rizoh (45, 39)
Digg-2	AmyVernon (103, 20), digits12 (0, 11), rocr69 (0, 13)
	Four20 (0, 13), ronzed (0, 3), guyro (0, 10)
	antdude (96, 11), basye (88, 6), openthink (0, 14)
	sirmcdiznuts (0, 11), netgeek06 (0, 20), warrior007 (0, 17)
	pirlok (0, 8), zen53 (0, 2), Morshade (0, 10)
Digg-3	hdar3415 (35, 31), zoomtechtv (0, 23), domfosnz (27, 27)
	EmitStop (68, 29), emberjohn (67, 9), oboy (0, 32)
	msaleem (0, 42), theeandrew (0, 21), MasterOfHyrule (0, 27)
	Surferess (0, 24), badwithcomputer (77, 53), tbhurst (0, 28)
	digits12 (76, 24), Visin (0, 7), kelvlam (3, 10)

Experiments

- Result: Representative network on a sub graph in Digg-2 Network



OUTLINE

- Background
- Challenge
- Unsupervised case 1
  - Representative user finding
- Unsupervised case 2
  - Community discovery
- Experiments
- Supervised case
  - Modeling information diffusion in social network

Modeling information diffusion in social network

- Supervised model
- Bridging the actual value (label) with the variable.
- More variables to come?
  - Learning the weights

Thanks