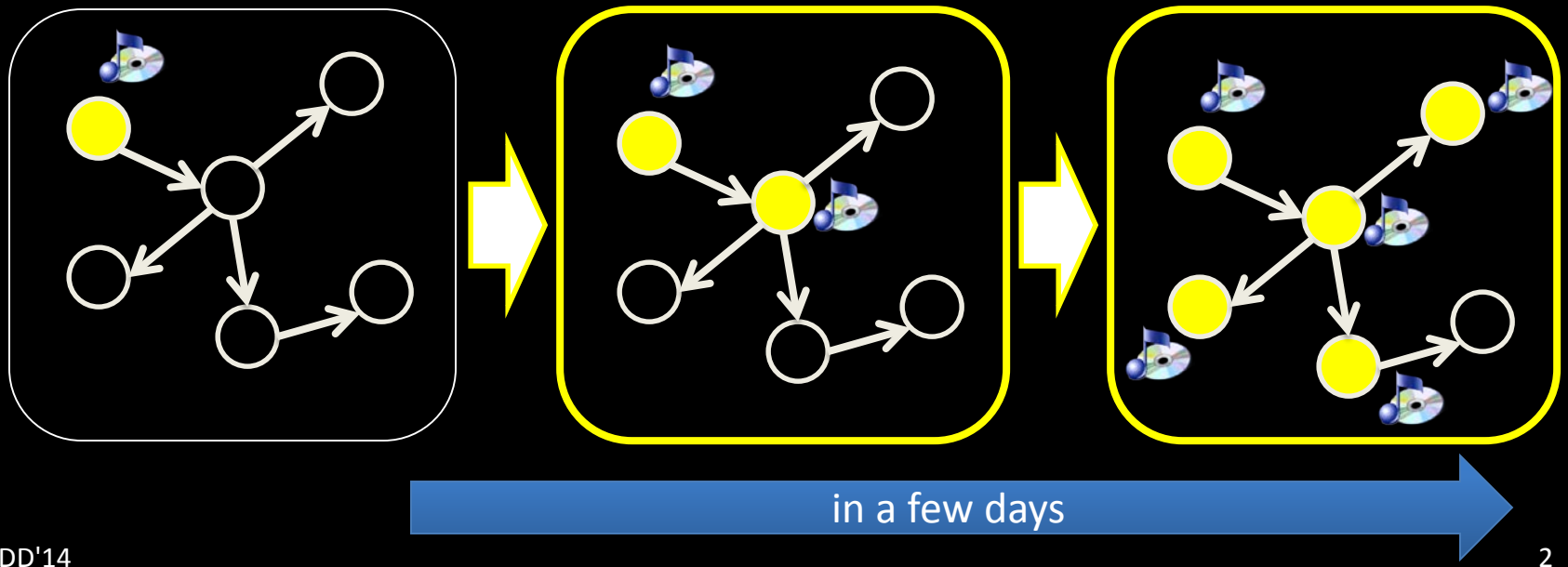
Probabilistic Latent Network Visualization: Inferring and Embedding Diffusion Networks

Takeshi Kurashima, Tomoharu Iwata,
Noriko Takaya, and Hiroshi Sawada

NTT Corporation, Japan

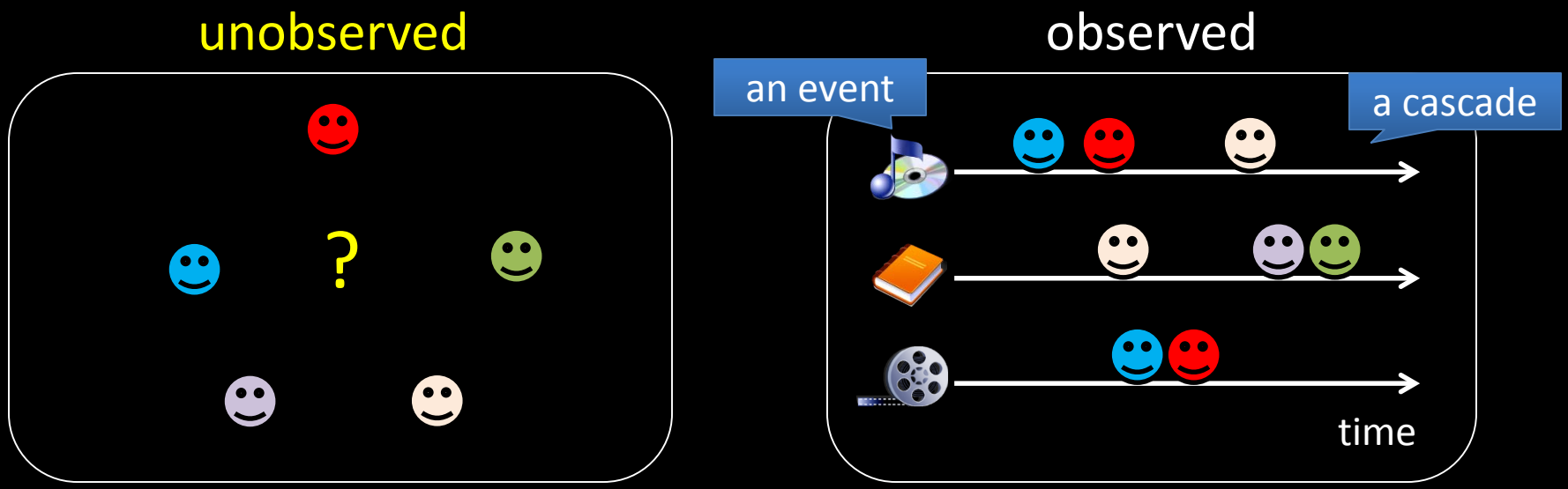
Introduction

- The diffusion of information, rumors, and diseases are assumed to be probabilistic processes over networks
- Understanding the mechanism (the network structure) that causes the diffusion helps to predict future events



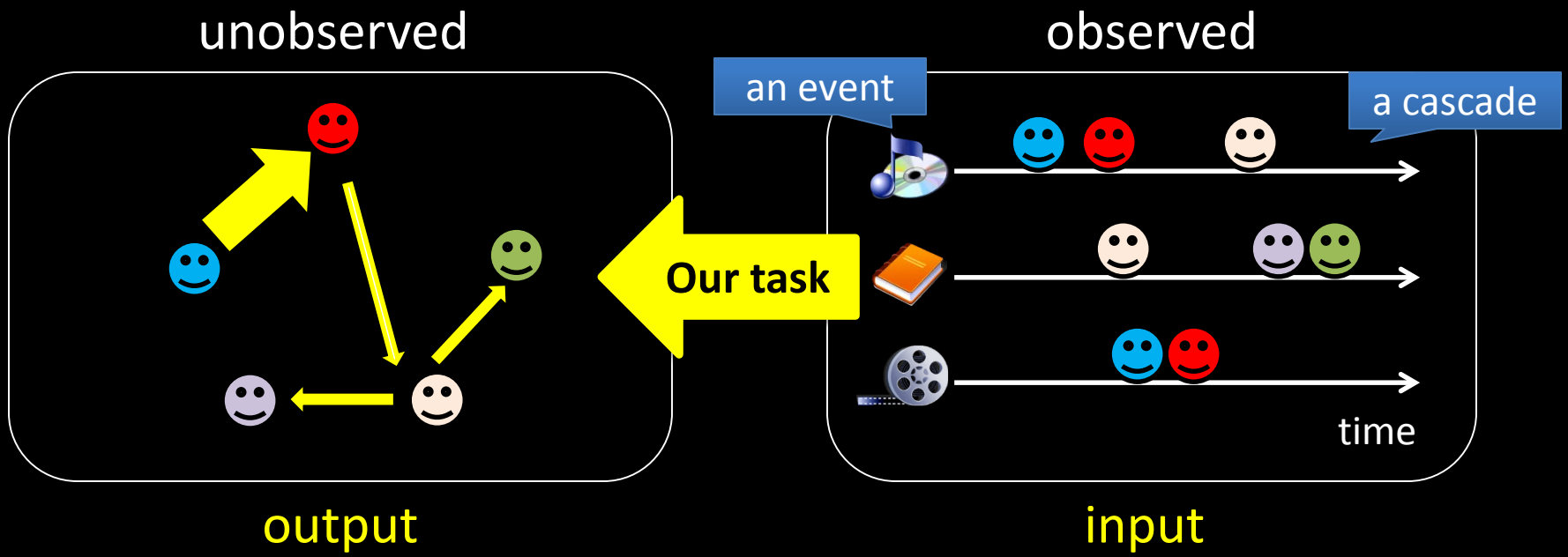
Hidden Network Structure

- Observation recorded only when a node mentions information, makes a decision, or becomes infected



Hidden Network Structure

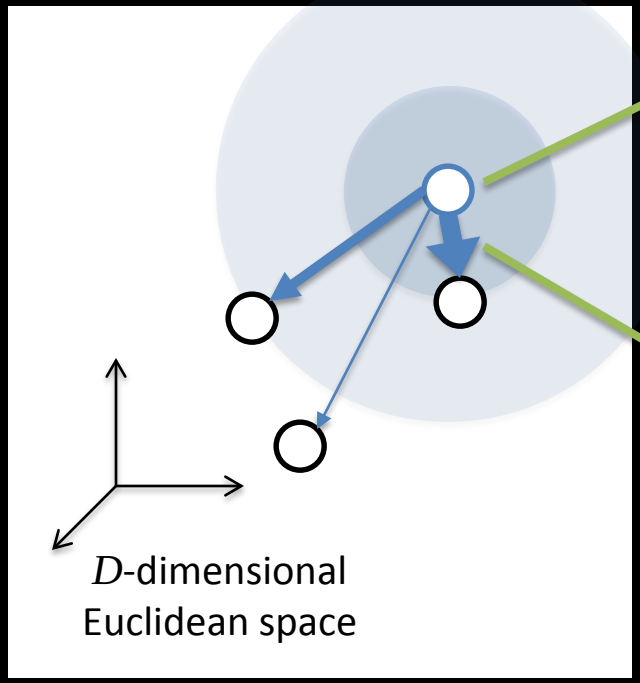
- Observation recorded only when a node mentions information, makes a decision, or becomes infected



Inferring diffusion networks based on cascade data

Proposed Model

- Key-feature: Inferring the diffusion network by embedding it into a low-dimensional Euclidean space.
- Our model learns the latent coordinates of nodes that best explain the observed cascade data



Each node n has latent coordinates $\boldsymbol{x}_n$ in the D-dimensional latent space

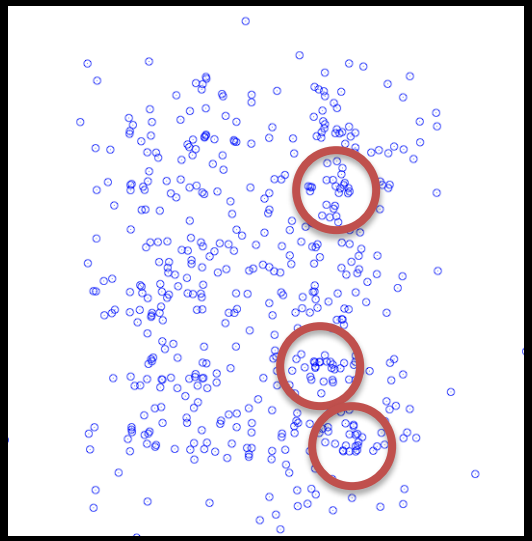
$$\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nD})$$

Diffusion is more likely to occur between nodes that are placed close together

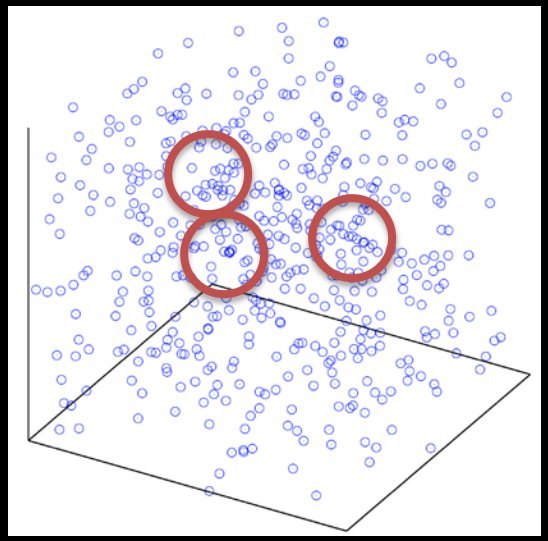
e.g.) an outbreak of influenza
-> geographical closeness between people

Advantage 1: Network Visualization

- *Influence preserving principle: each node attempts to place its influential nodes relatively closer than non-influential ones*



2-dimensional visualization



3-dimensional visualization

From the initial sight of the layout, we can identify *communities* wherein the nodes strongly influence each other

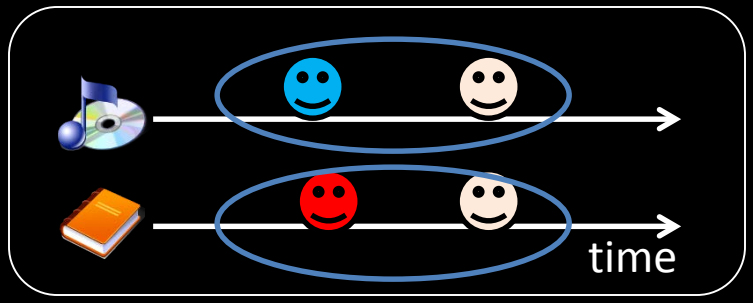
Advantage 2: Network Inference

- High accuracy even when many cascades remain hidden
  e.g. the diffusion process of new or rare information and disease
- The number of parameters to be estimated is small

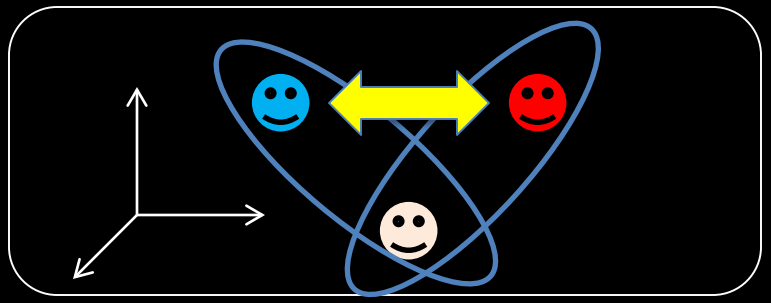
Existing model	Proposed model
$N \times N >> D \times N$	

N: number of nodes
D: dimensionality of the latent space

- The short distance between two nodes in the latent space implies the presence of influence-relation



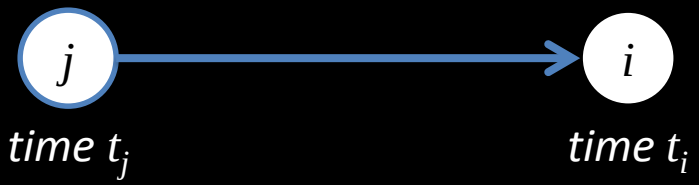
Cascades (input)



Node coordinates in the latent space (output)

Preliminary: Survival Analysis

- *The transmission function*:
  The likelihood of an event happening to i at time t_i
  given that the same event has already happened to j at time t_j



$$f(t_i|t_j) = f(\Delta_{ji})$$
$$\Delta_{ji} = t_i - t_j$$

- *The survival function*: The probability that i is NOT infected by j before time t_i

$$S(t_i|t_j) = \int_{t_i}^{\infty} f(t)\, dt.$$

- *The hazard function*: The rate that i becomes infected by j after time t_i

$$h(t_i|t_j) = \frac{f(t_i|t_j)}{S(t_i|t_j)}$$

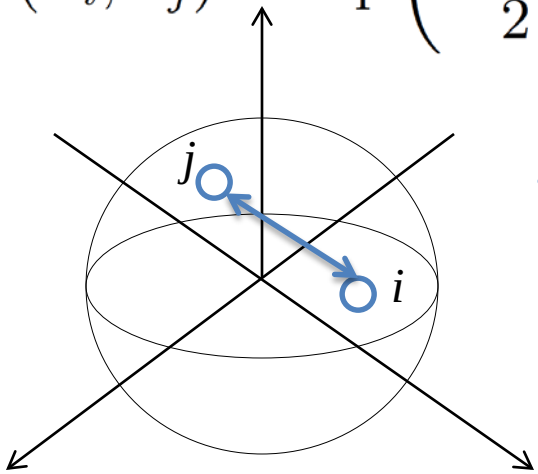
Proposed Model

The transmission function

$$f\left(\Delta_{ji}|\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\mu\alpha\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\left(\Delta_{ji}\right)^{\mu-1}\exp\left(-\alpha\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\left(\Delta_{ji}\right)^{\mu}\right)$$

The transmission rate: *how likely node j is to infect node i*

$$\alpha\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\exp\left(-\frac{\beta}{2}||\boldsymbol{x}_j-\boldsymbol{x}_i||^2\right)$$



e.g. D=3

The Euclidean distance in the D-dimensional latent space

Likelihood

- MAP (maximum a posteriori) estimation
  - The unknown parameters -> node coordinates

$$\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}, \text{ where } \boldsymbol{x}_n = (x_{n1}, ..., x_{nD})$$

- The negative log likelihood of parameters $\boldsymbol{X}$ for the given cascades, $\boldsymbol{C}$, with prior is as follows:

$$L(\boldsymbol{X}|\boldsymbol{C}) = \sum_i \sum_j \sum_{\{c|t_j^c < t_i^c\}} \exp\left(-\frac{\beta}{2}||\boldsymbol{x}_j - \boldsymbol{x}_i||^2\right) (\Delta_{ji}^c)^{\mu}$$

$$- \sum_i \sum_{\{c|t_i^c \leq T^c\}} \log \sum_{\{j|t_j^c \leq t_i^c\}} \mu \exp\left(-\frac{\beta}{2}||\boldsymbol{x}_j - \boldsymbol{x}_i||^2\right) (\Delta_{ji}^c)^{\mu-1}$$

$$- \sum_i \log\left(\left(\frac{\gamma}{2\pi}\right)^{\frac{D}{2}} \exp\left(-\frac{\gamma}{2}||\boldsymbol{x}_n||^2\right)\right),$$

Using a Gaussian prior with zero mean and spherical covariance for $\boldsymbol{x_n}$

Inference

- The negative log likelihood can be minimized through the use of the gradient-based numerical optimization method

The gradient of $\boldsymbol{x}_n$

$$\frac{\partial L}{\partial \boldsymbol{x}_n} = \sum_j \sum_{\{c|t_n^c \neq t_j^c\}} \beta\left(\boldsymbol{x}_j - \boldsymbol{x}_n\right) \exp\left(-\frac{\beta}{2}||\boldsymbol{x}_j - \boldsymbol{x}_n||^2\right) \left(\Delta_{jn}^c\right)^\mu$$

$$- \sum_{\{c|t_n^c \leq T^c\}} \sum_{\{j|t_j^c < t_n^c\}} \beta\left(\boldsymbol{x}_j - \boldsymbol{x}_n\right) \boxed{P\left(n|j\right)}$$

The conditional probability of infection from node j to node n

$$- \sum_j \sum_{\{c|t_j^c \leq T^c\}} \beta\left(\boldsymbol{x}_j - \boldsymbol{x}_n\right) P\left(j|n\right) + \gamma \boldsymbol{x}_n,$$

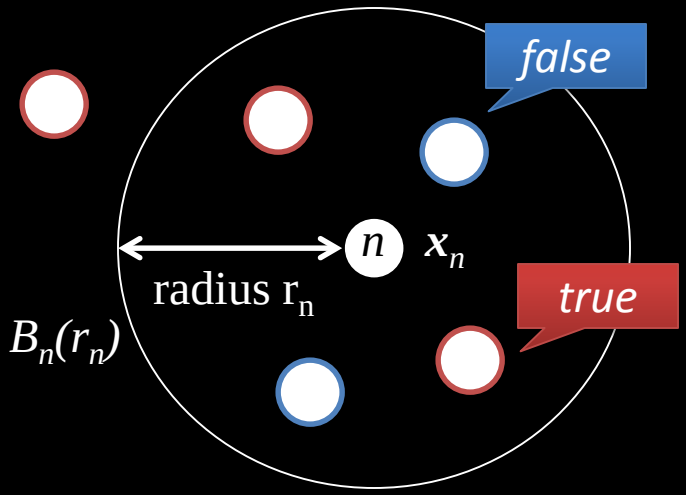
Experiments: Data

- The information diffusion occurring in Web
  - When a news article or blog used a keyword
  - MemeTracker dataset of [Rodriguez et al.,WSDM'13]
- Eight types of cascade data
  - "iPhone", "Jobs", "Baseball", "Basketball"
    "Earthquake", "Fukushima", "Sept.11", "Syria"

Exp.1: Visualization Performance

- The visualization space learned from training data could predict the influence-relation in the test data
  - The influence-relation between nodes is *true* if test data has an event wherein both are infected

Evaluation metric: *F-measure*

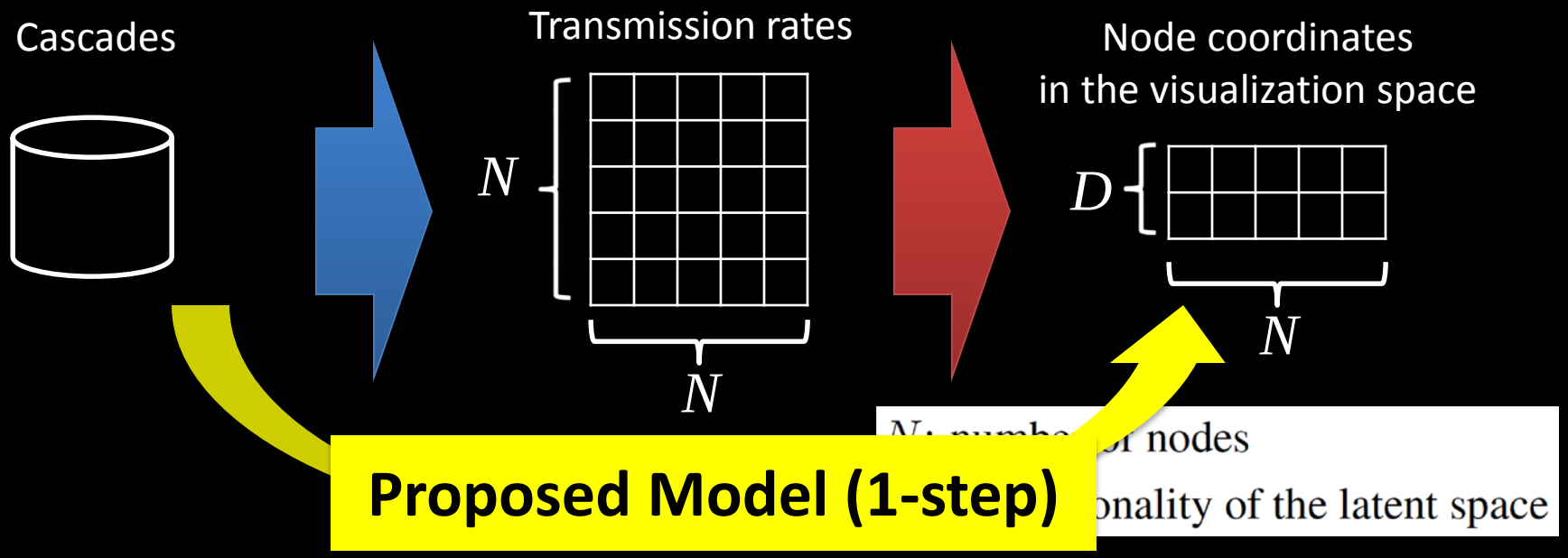


The precision for node n:
the fraction of nodes that are contained in $B_n(r_n)$ that have influence-relation to node n

The recall for node n:
the fraction of nodes that have influence-relations that are successfully contained in $B_n(r_n)$

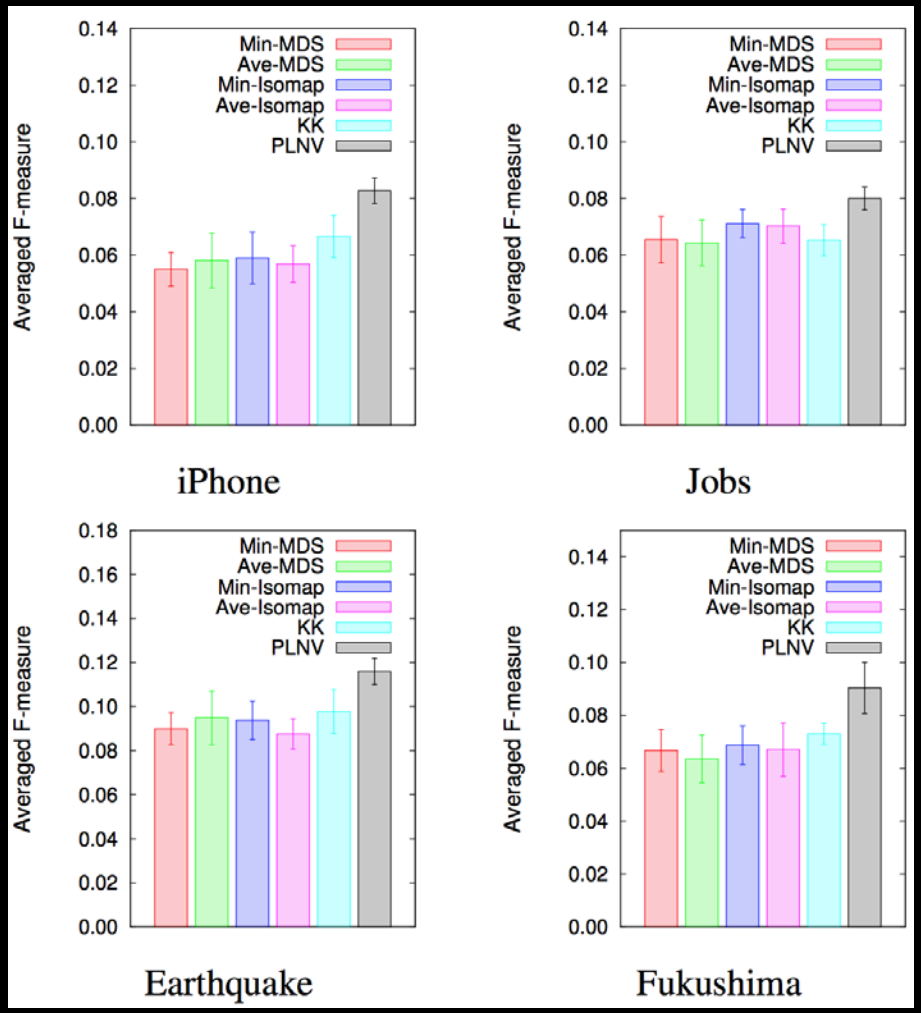
Baseline Method

- 2-step process
  1. Inferring the diffusion network using NETRATE [ICML'11]
  2. Embedding it into 2-, and 3-dimensional visualization space using MDS, Isomap, or KK-spring method

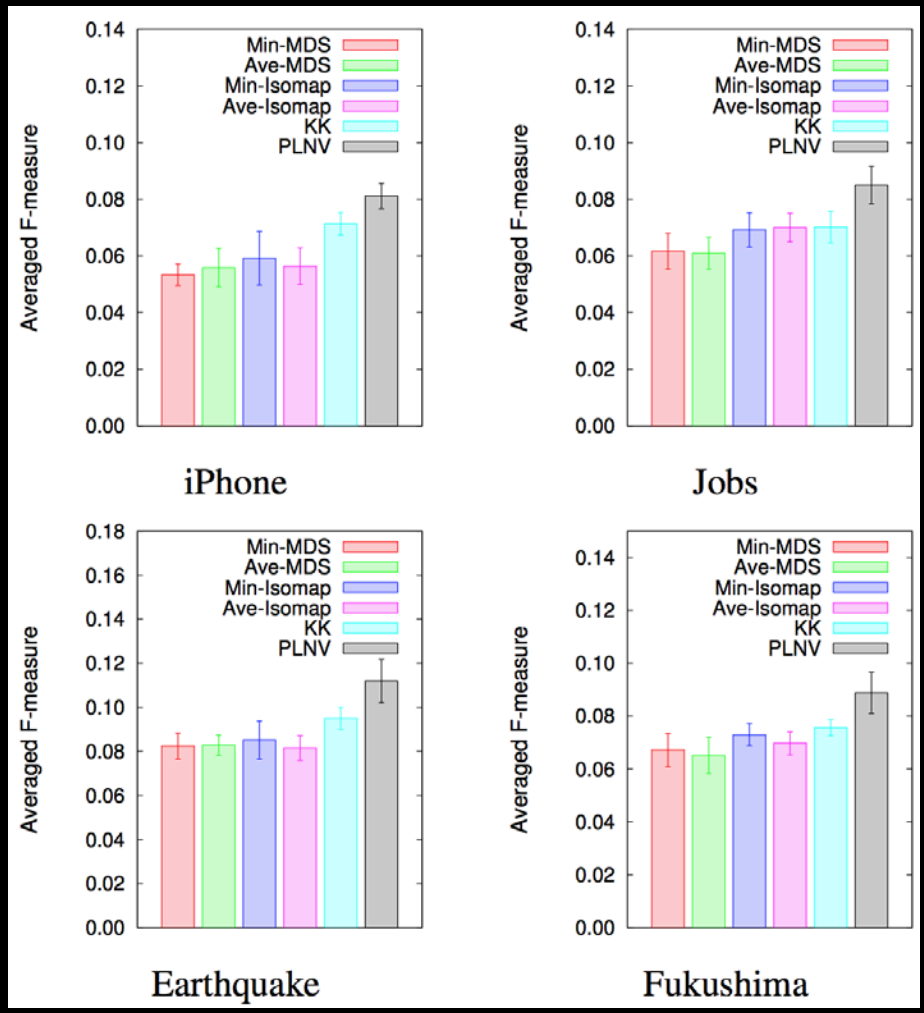


Exp.1: Visualization Performance

PLNV satisfies the *influence-preserving-principle* in the visualization space



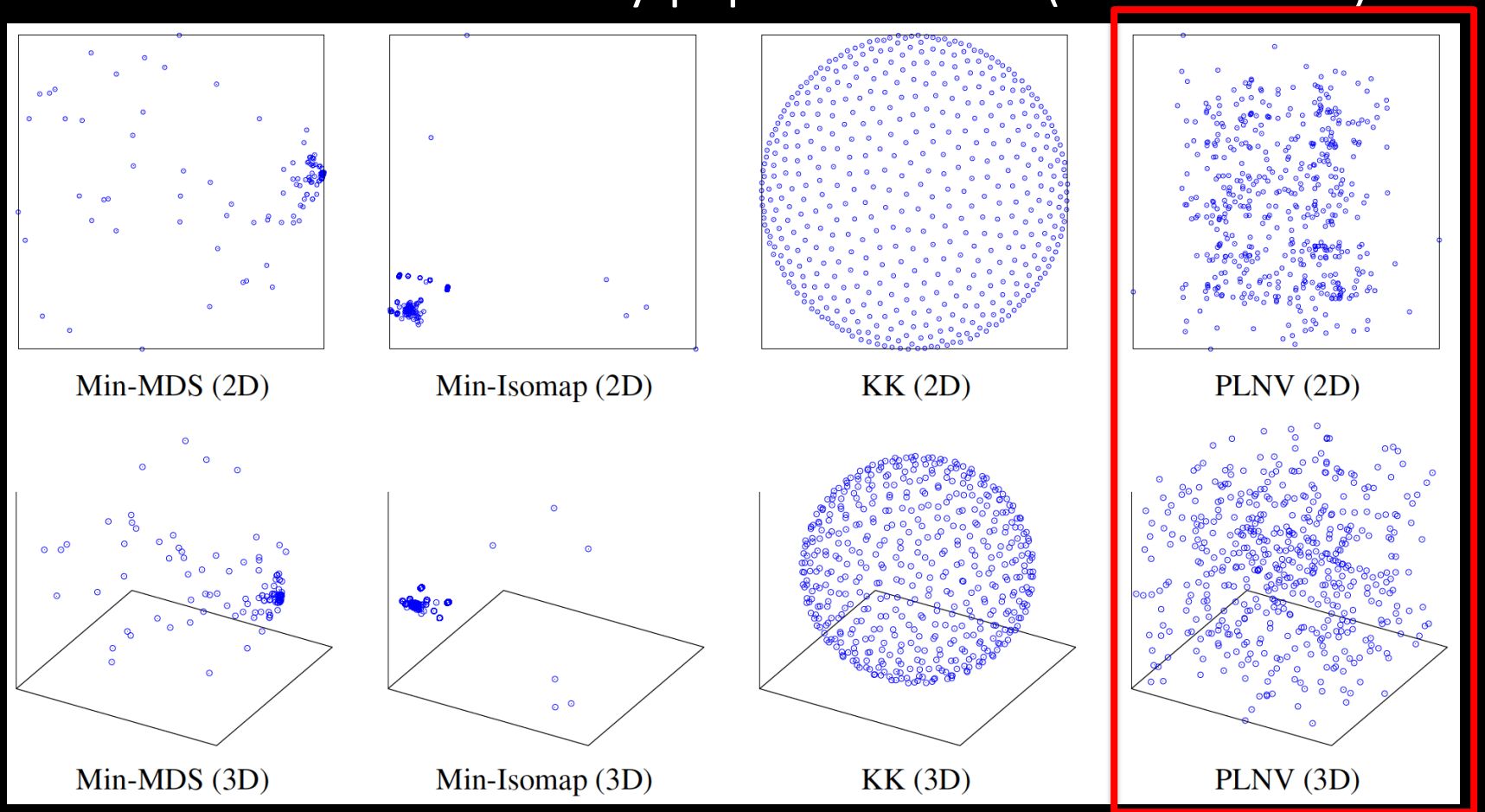
2-dimensional



3-dimensional

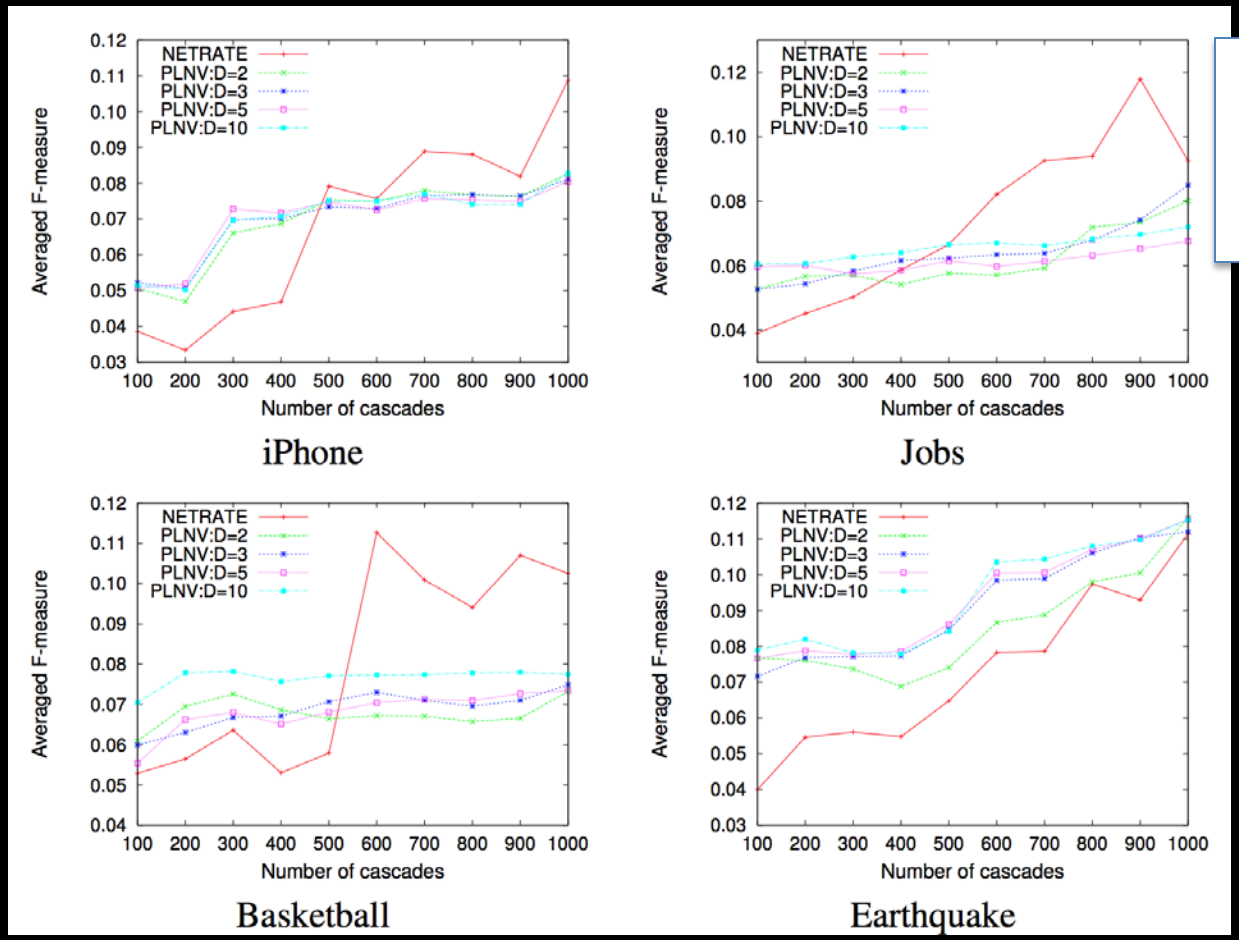
Comparisons of 2-, and 3-dimensional Visualization

PLNV avoids the node collapse problem,
and exhibits several densely-populated areas (communities)



Exp.2: Network Inference Performance

PLNV is appropriate for solving the network inference problem with sparse cascade data (# cascades < 300)



RED: NETRATE(baseline)
OTHERS: PLNV(proposed)
D=2,3,5, and 10

Conclusion

- We propose a probabilistic model for inferring & visualizing diffusion networks
  - Suggests the network layout that satisfies the *influence-preserving principle*
  - Accurately infers the diffusion network when the number of cascades is relatively small
- Future work
  - Other types of cascade data (e.g. infectious disease)
  - Topic dependent diffusion