Cascading Outbreak Prediction in Networks: A Data-Driven Approach

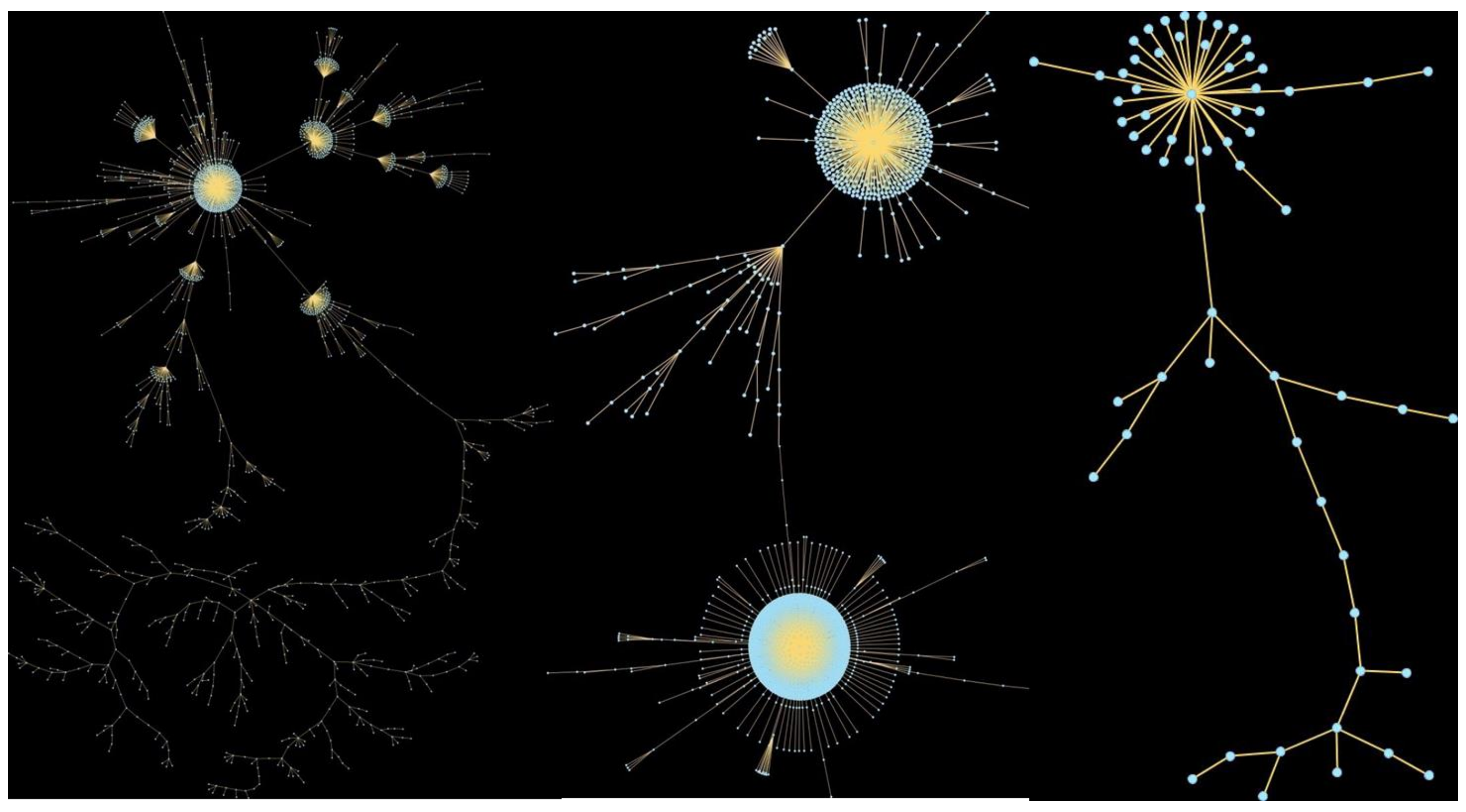
Peng Cui[1], Shifei Jin[1], Linyun Yu[1], Fei Wang[2], Wenwu Zhu[1], Shiqiang Yang[1]

[1]Department of Computer Science, Tsinghua University. Beijing, China
[2]IBM Watson Research Center, Yorktown Heights. NY, U.S.

cuip@tsinghua.edu.cn, shifei.jin@gmail.com, linyun.yu.08@gmail.com, fwang@us.ibm.com,
wwzhu@tsinghua.edu.cn, yangshq@tsinghua.edu.cn

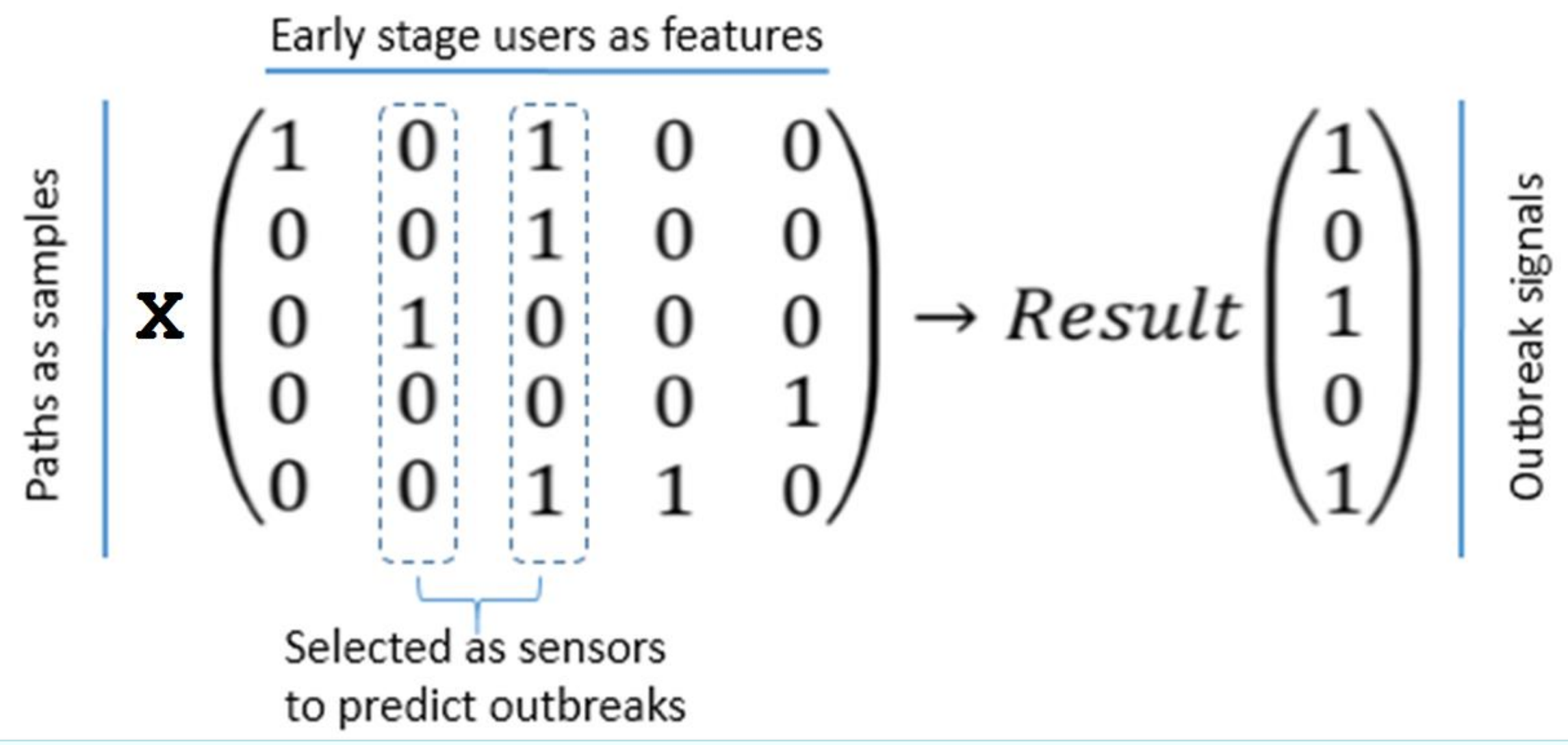
Motivation

- ➢ Cascades are ubiquitous in various network environments such as epidemic networks, traffic networks, water distribution networks and social networks.
- ➢ The outbreaks of cascades will often bring bad or even devastating effects.
- ➢ How to accurately predict the cascading outbreaks in early stage?



OSLOR: Orthogonal Sparse LOgistic Regression

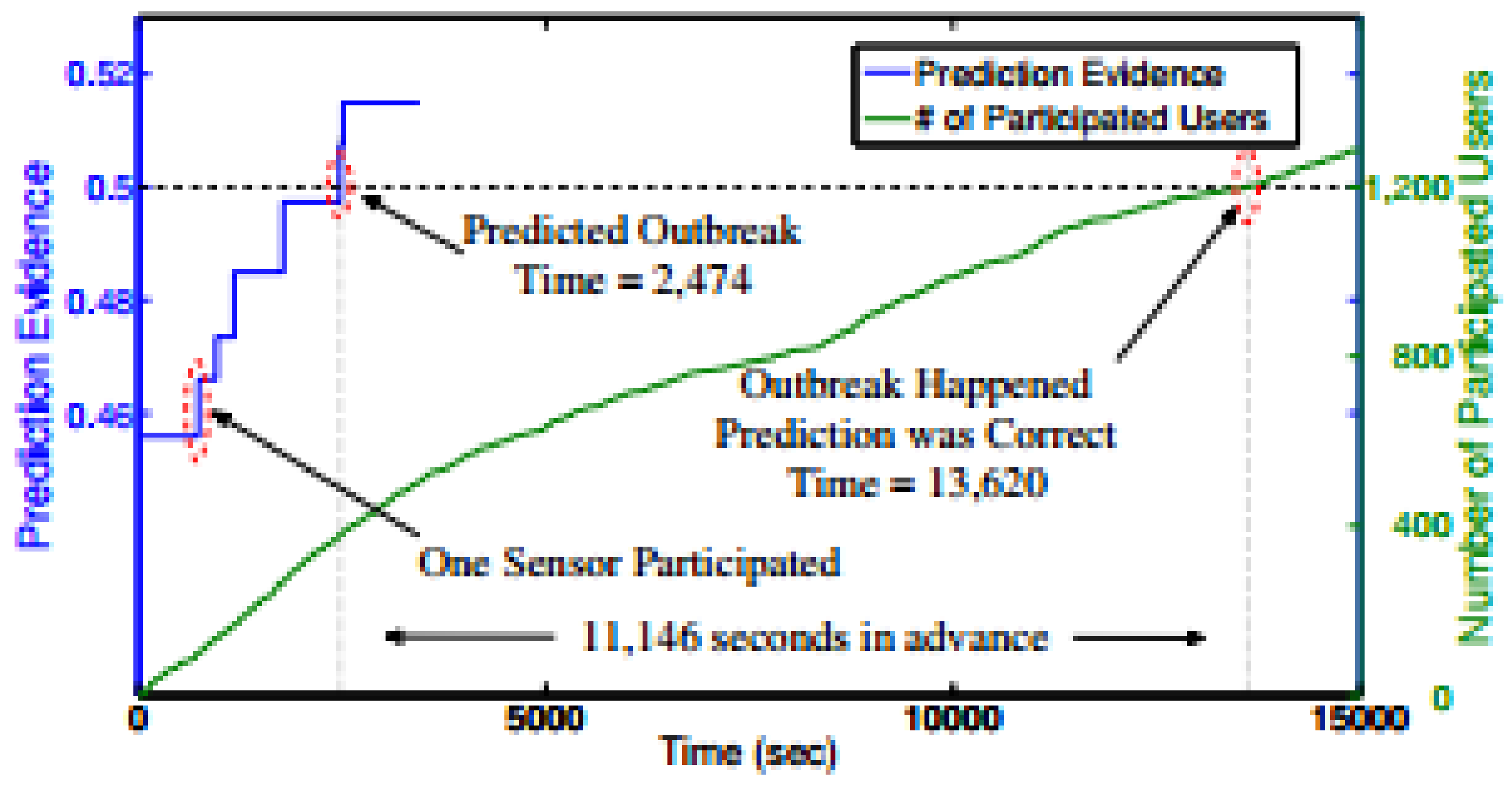
- ➢ Basic Hypothesis: User behaviors cause outbreaks
- ➢ To discover the users whose behaviors are highly correlated with the outbreaks from large data, and select them as sensors to monitor.



$$\mathbf{X}\begin{pmatrix}1&0&1&0&0\\0&0&1&0&0\\0&1&0&0&0\\0&0&0&0&1\\0&0&1&1&0\end{pmatrix}\rightarrow Result\begin{pmatrix}1\\0\\1\\0\\1\end{pmatrix}$$

Decision function:

$$h(\mathbf{X}_{i,}^{t}) = sigmoid(\theta_0 + \mathbf{X}_{i,}^{t}\theta) = \frac{1}{1+\exp(-\theta_0 - \mathbf{X}_{i,}^{t}\theta)}$$



early prediction of cascading outbreaks by OSLOR.

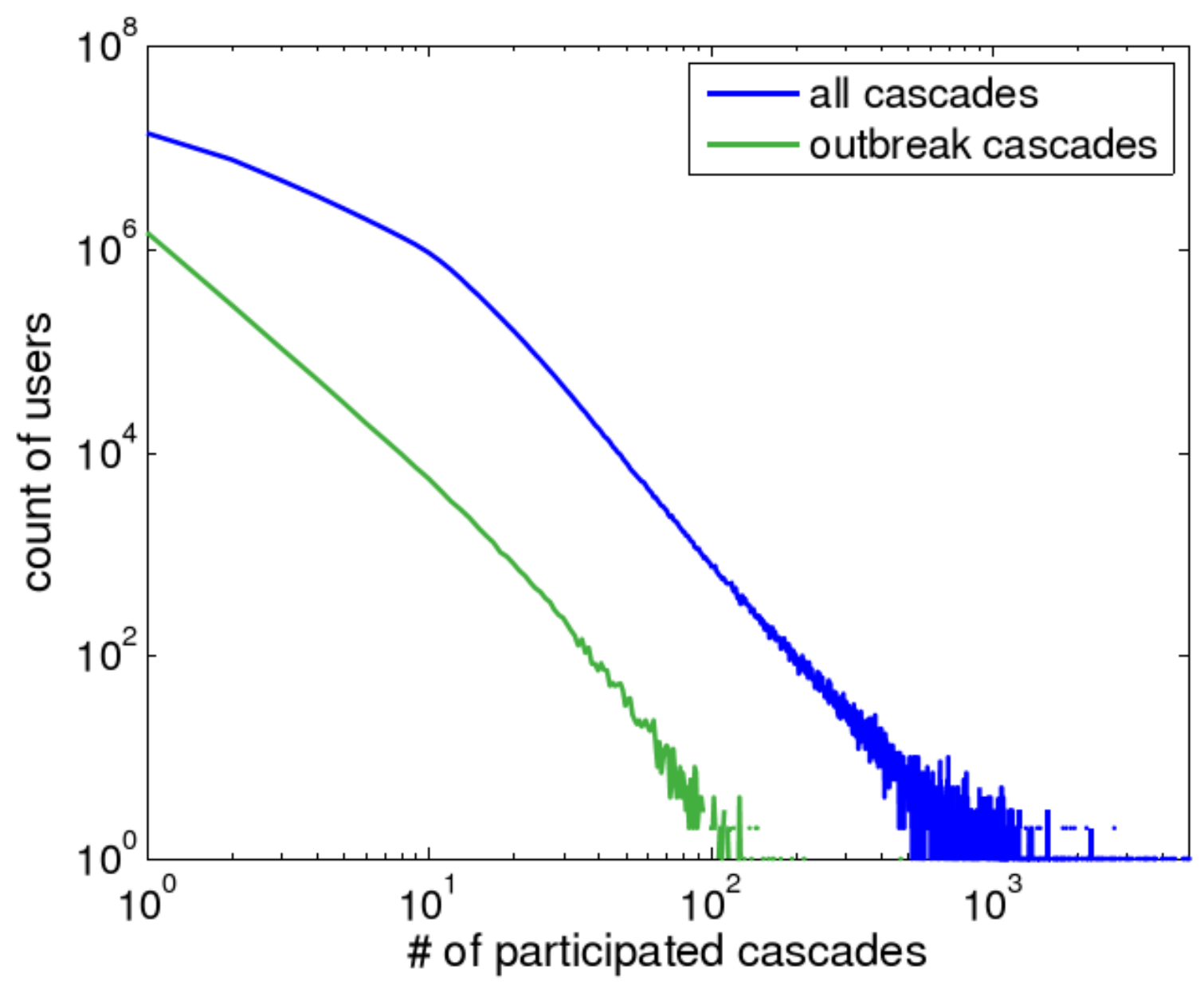
Minimizing target:

$$F(\theta) = T_1(\theta) + T_2(\theta) + T_3(\theta)$$

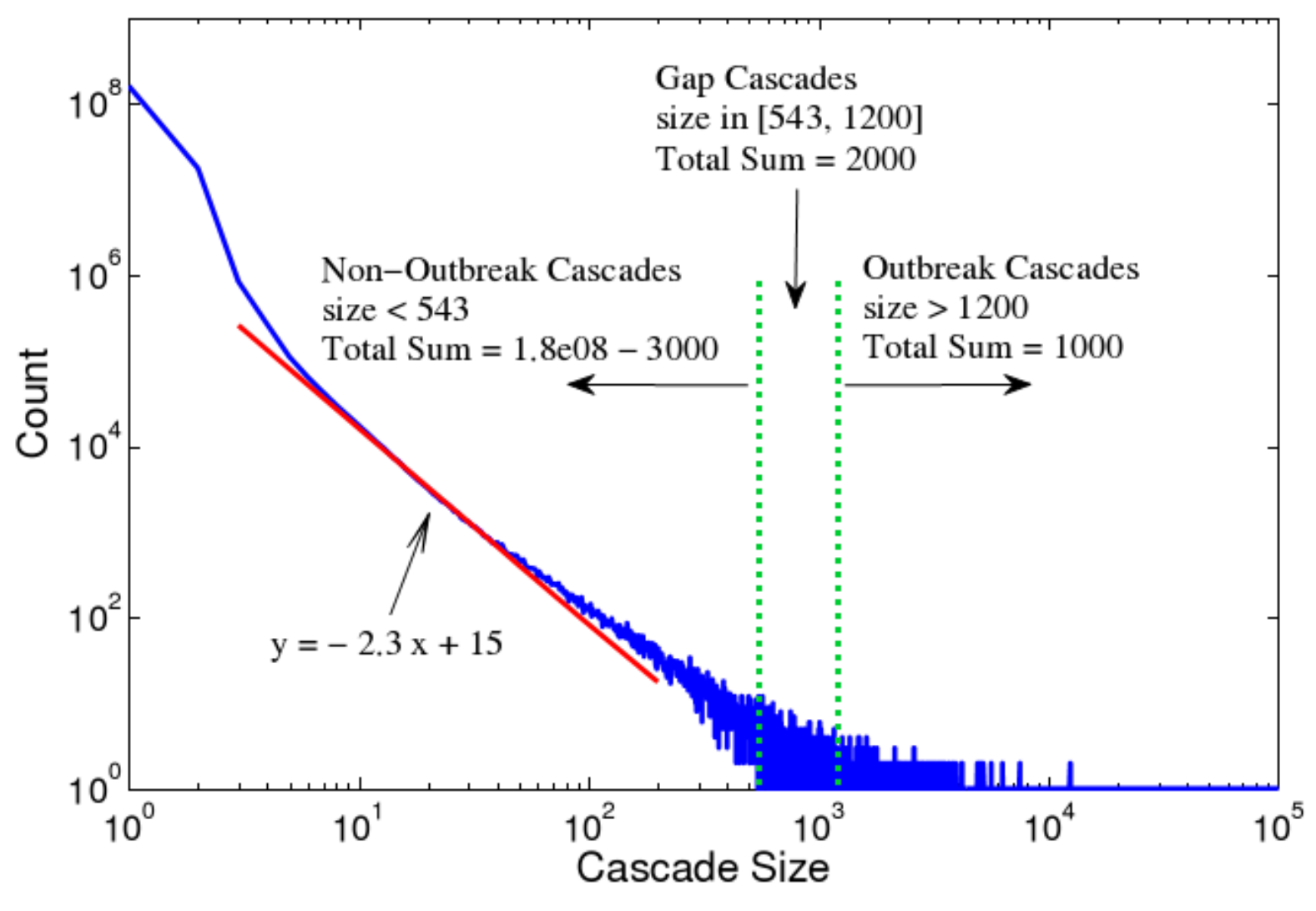
$T_1(\theta) = -\log L(\theta)$ $T_1(\theta)$: the target of logistic regression.

$T_2(\theta) = \frac{\beta}{4}\sum_{i,j}(\theta_i \mathbf{X}_{\cdot i}^{\top}\mathbf{X}_{\cdot j}\theta_j)^2$ $T_2(\theta)$: The powerful users should have minimum redundancy.

$T_3(\theta) = \gamma\|\theta\|_1$ $T_3(\theta)$: The powerful users should be limited.

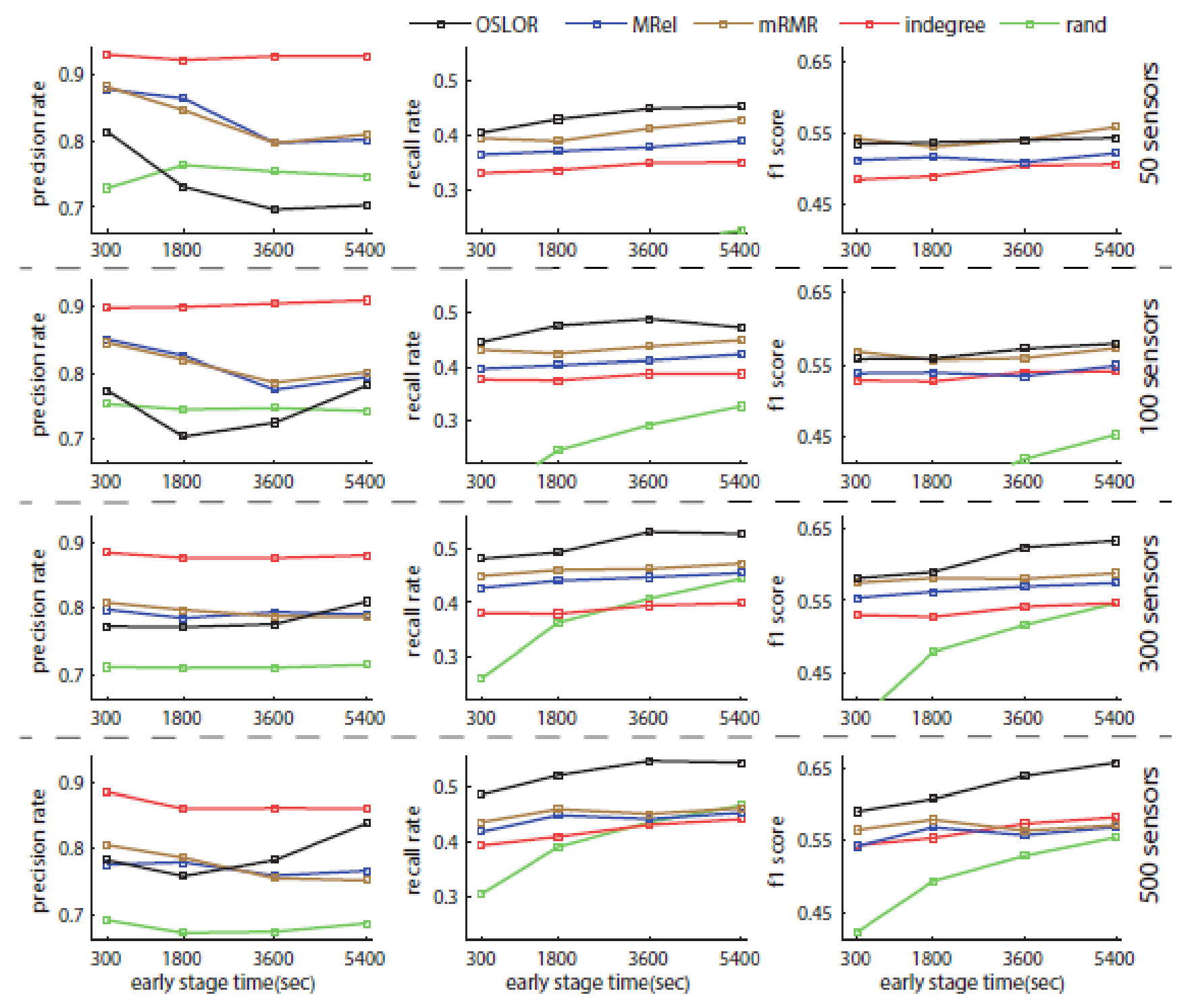


The distribution of the quantity of (outbreak) cascades that users participate.

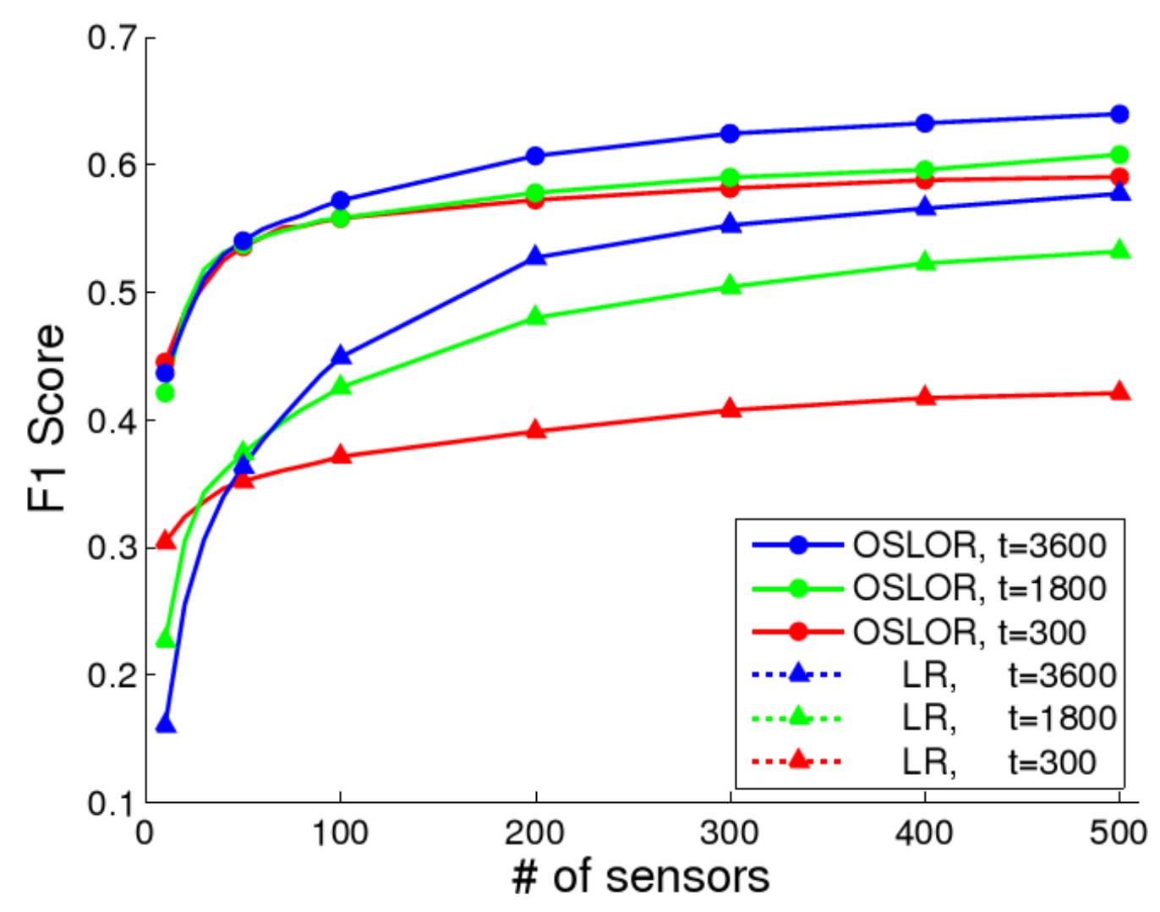


The distribution of cascade size

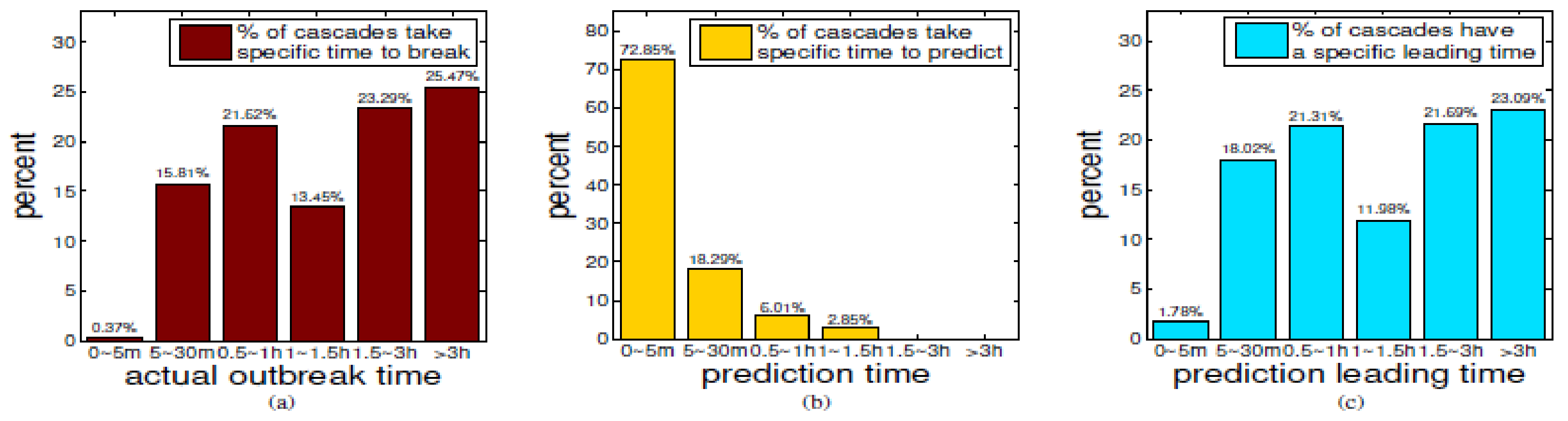
Prediction Performance



Prediction results of different methods with different early stage time.



Comparison of OSLOR and logistic regression.



We only need 5 mins to predict the information outbreaks!

- ➢Our approach performs best
- ➢Data driven approaches outperforms topology-based approaches

- ➢Big nodes' participation will cause outbreaks in most cases
- ➢Only a part of outbreaks are caused by big nodes

Paper ID: 0554