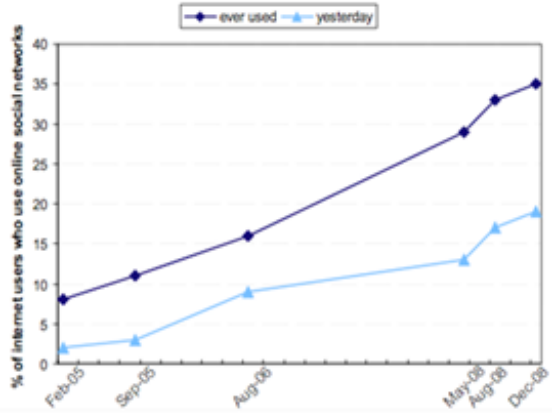
복잡계 망의 정보 흐름 최적화 연구

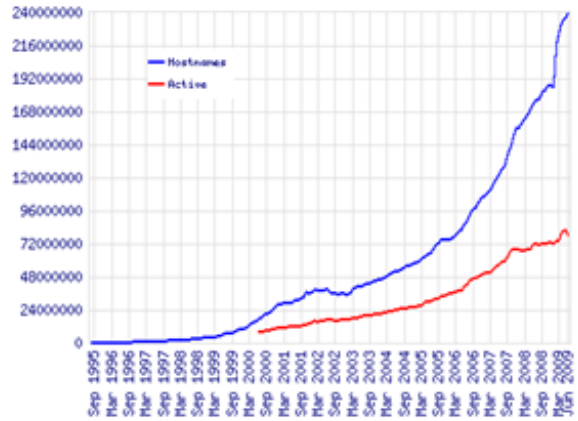
정교민
Applied Algorithm Lab
전산학과, KAIST

복잡계 망(Complex Network)

- 구성 요소들이 긴밀하게 연결되어 상호 작용하는 네트워크 시스템
- 전산, 물리, 수학, 경제학, 사회학 등에서 활발히 연구되는 융합적 학문
- WWW, 프로그램 구조망, 소셜 네트워크, 바이러스 전파망 등
  - 최근 가파른 성장세를 보임



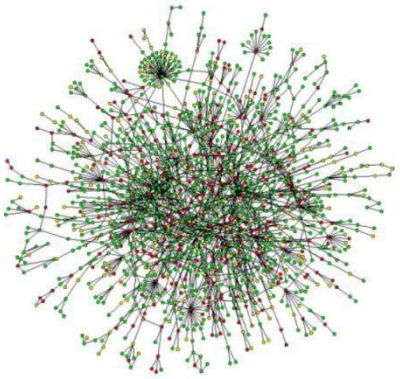
기간별 소셜 네트워크 성장세



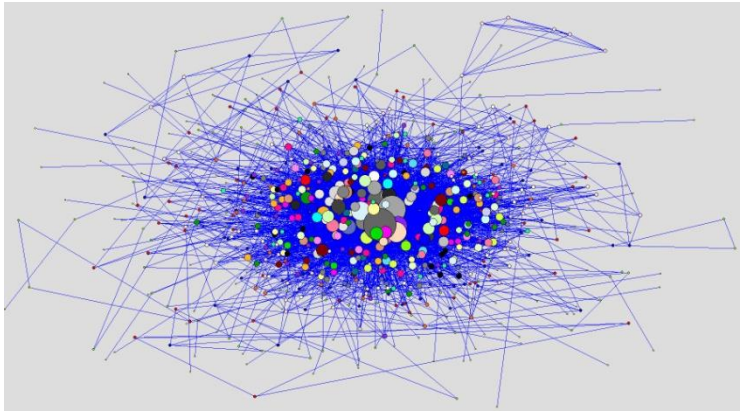
WWW의 연도별 성장세

복잡계 망의 성질

인터넷 망
AT&T and Lumeta



이스트 단백질 네트워크
P. Uetz 공저, Nature 2000



쿼리 네트워크
CLAIR(University of Michigan)

- 좁은 세상(small world) 구조
- 불균등한 위상 구조 (power law distribution)
- 유사한 개체들 간에 긴밀히 연결되어 있는 커뮤니티 구조 (community structure)

연구 동향

모델링

- 복잡계망의 척도없는 모델 (Albert et al., 1999)
- 정보 흐름 모델 (Independent Cascade model, Threshold model)

최적화 및 머신 러닝

- 망에서의 정보 흐름 최적화 (Kempe et al., 2003, Leskovec et al., 2007)
- 노드 간의 연결 추천 (Liben-Nowell and Kleinberg, 2003), 노드 분류 (Zhu et al., 2007) 문제를 머신 러닝 기법을 사용해 접근
- 모듈화 함수 최적화를 통한 집단 검출 문제 (Newman and Girvan, 2004)

분산화

- 정보 전달을 분산적으로 처리하는 가십 알고리즘 (Boyd et al., 2006)
- 망에서의 지역 연산을 통해 전역 최적으로의 근사 (Jung et al., 2009)

복잡계 망 정보 흐름 최적화 연구

1. 복잡계 망 구조 분석을 통한 정보 흐름 특성 이해

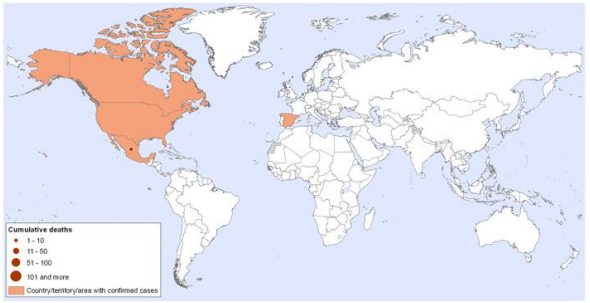
2. 머신 러닝 기법을 통한 정보 흐름 최적화 문제 해결

3. 복잡계 망의 동적 · 질적 변화에 유연한 분산 알고리즘 개발

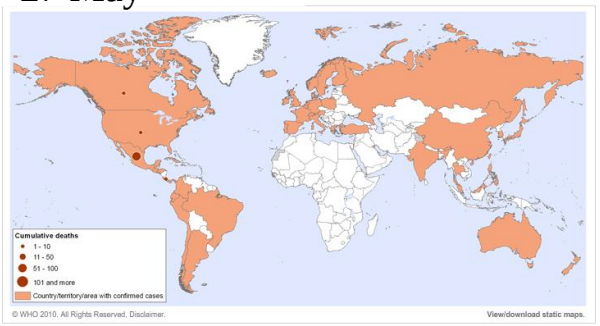
1. 정보 확산의 특성 이해

- 정보 확산이 급격히 일어나는 **상전이 현상**(phase transition)의 조건 규명
  - **전염병, 컴퓨터 바이러스** 확산 패턴, **혁신·정치적 견해**의 전파 패턴

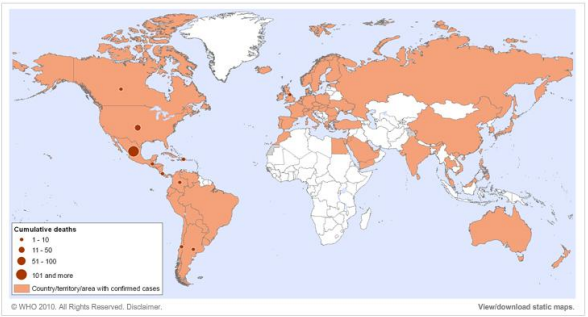
27 April



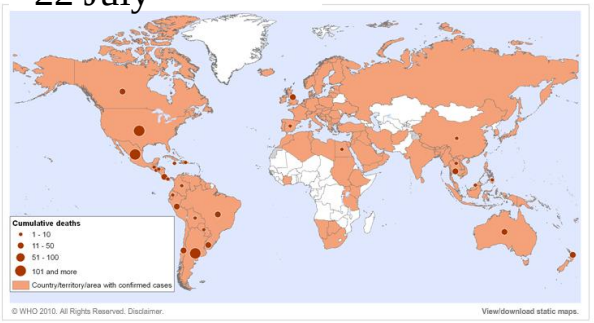
27 May



17 June



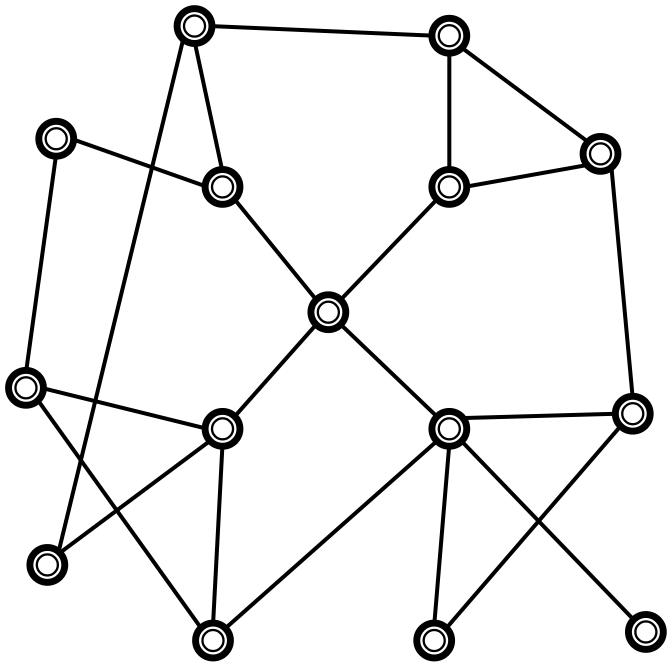
22 July



〈2009년 월별 H1N1 확산〉

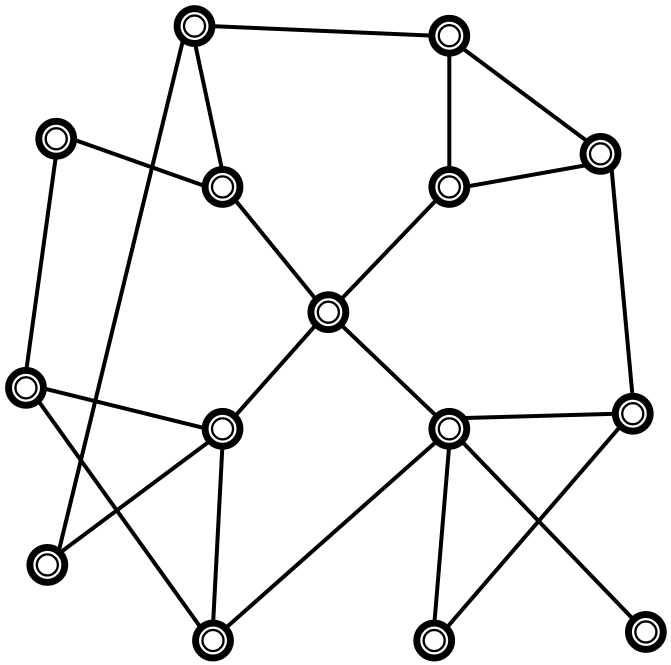
1. 정보 확산의 특성 이해

- 정보 확산이 급격히 일어나는 상전이 현상(phase transition)의 조건 규명
  - 전염병, 컴퓨터 바이러스 확산 패턴, 혁신·정치적 견해의 전파 패턴



1. 정보 확산의 특성 이해

- 정보 확산이 급격히 일어나는 상전이 현상(phase transition)의 조건 규명
  - 전염병, 컴퓨터 바이러스 확산 패턴, 혁신·정치적 견해의 전파 패턴



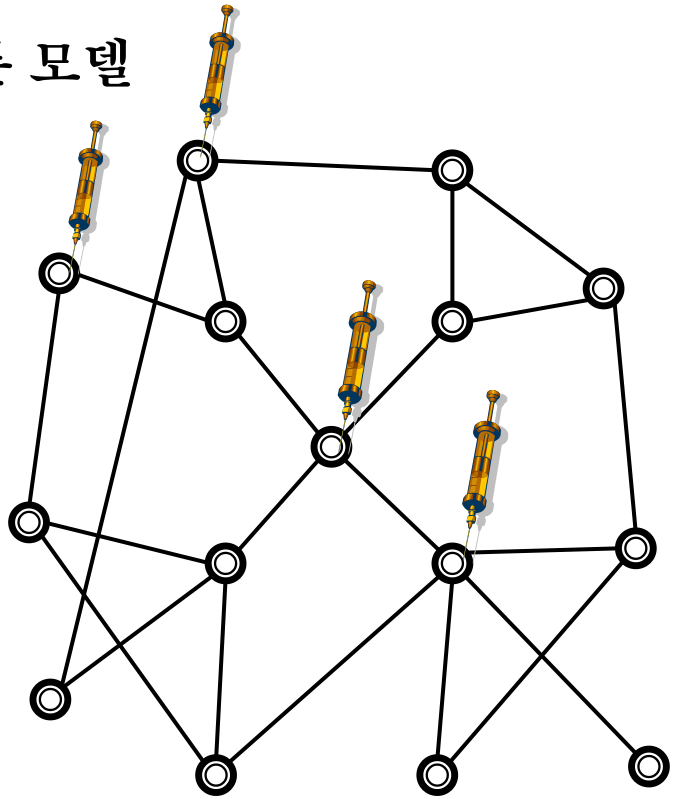
2. 정보 흐름 최적화

- 한정된 자원의 효율적인 분배
  - 정보 확산 촉진 및 억제
  - 이상 현상 감시
  - 커널 기법 기반의 머신 러닝, 통계적 추론 모델

- 온라인 마케팅
  - 온라인 상의 복잡계 망에서 입소문 현상을 이용한 마케팅
  - 한정된 재화로 광고를 할 때, 대상을 선정하는 문제

2. 정보 흐름 최적화

- 한정된 자원의 효율적인 분배
  - 정보 확산 촉진 및 억제
  - 이상 현상 감지
  - 커널 기법 기반의 머신 러닝, 통계적 추론 모델
- 백신 분배 문제



3. 분산 알고리즘 개발

- 빠르게 진화하는 large scale 망에도 적용 가능해야 함
- 분산 연산 기반 복잡계 망 이상 현상 감지 알고리즘 개발
- 분산 시스템 하에서의 최적화 알고리즘 구현

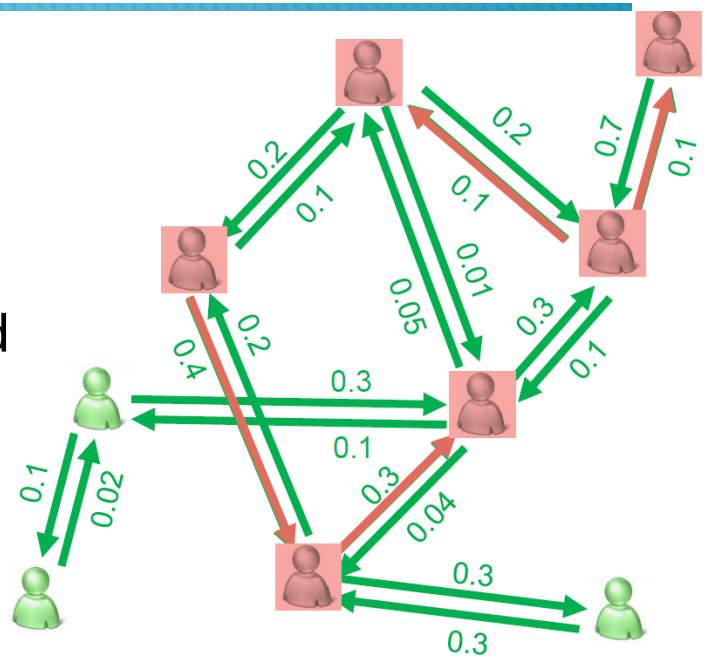
정보 흐름 분석 연구

- Traditionally the diffusion of innovation is studied in Sociology
  - Adoption of hybrid corn (Ryan and Gross, 1943)
  - Diffusion of innovations among physicians (Coleman et al., 1957)
  - Innovation decision process theory (Rogers, 1962)

- Lots of mechanisms have been investigated
  - Independent Cascade model
  - Linear threshold model

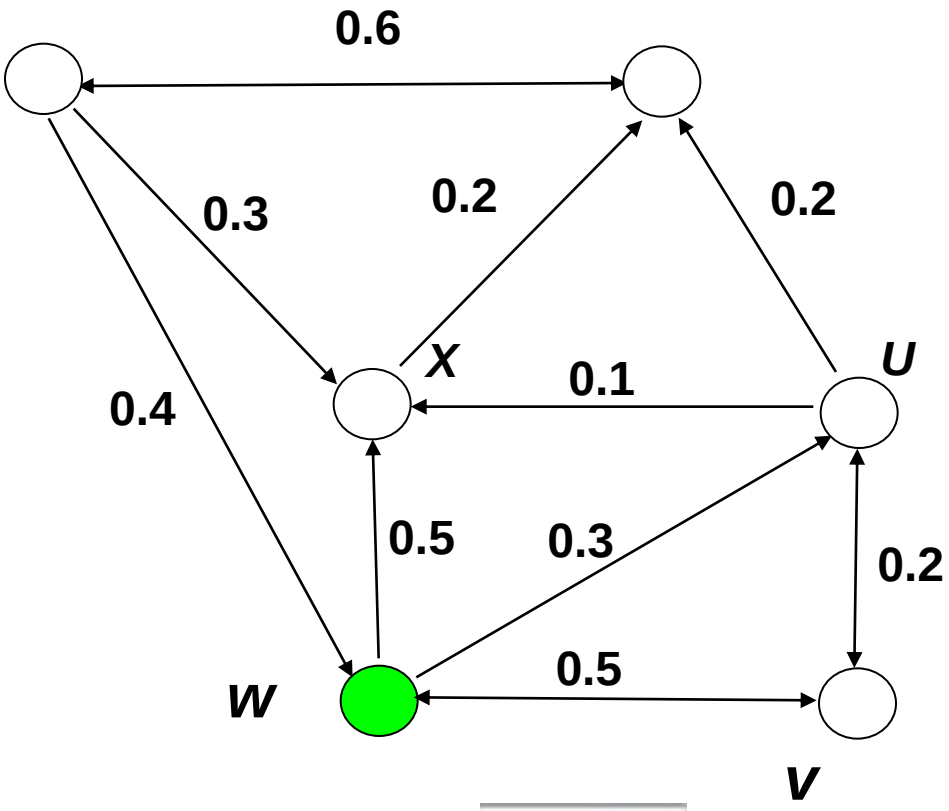
Independent cascade model

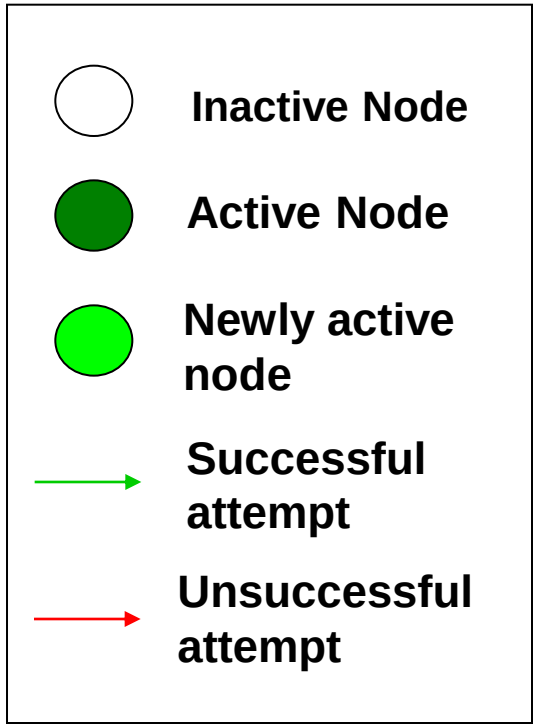
- Independent cascade model (IC)
  - Each node has active or inactive state
  - Initially some seed nodes are activated
  - At each step, each newly activated node u tries to activate its neighbor v with probability P_{uv}



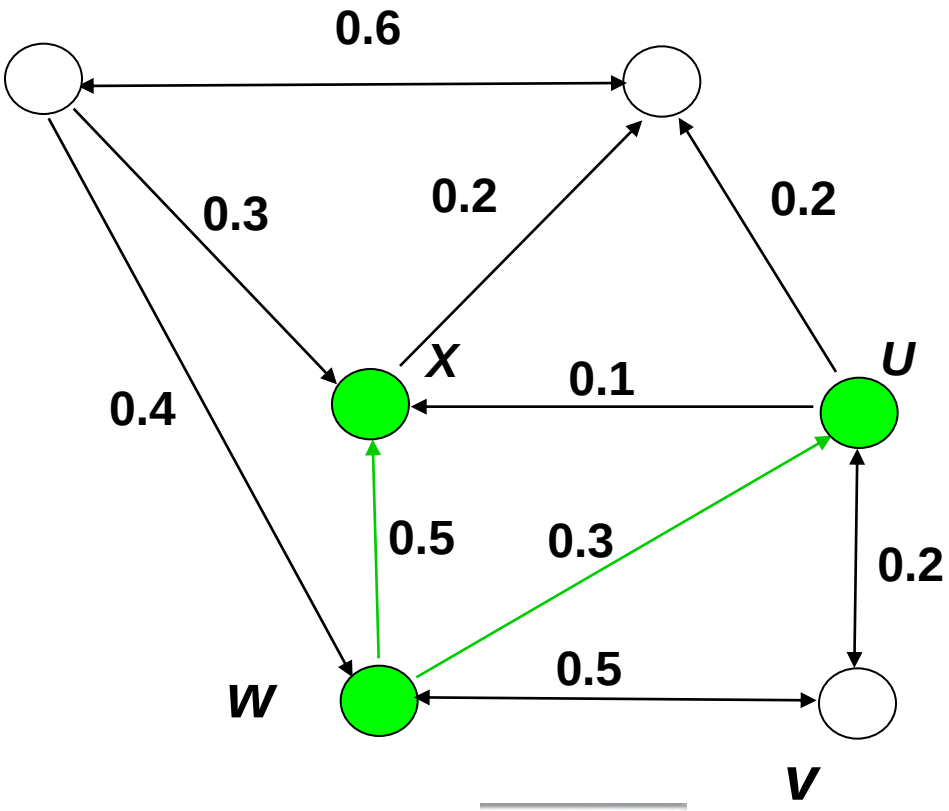
- This process explains information diffusion in complex networks.
  - Facebook, Twitter retweet, information spreading in the blog space, epidemic spreading etc

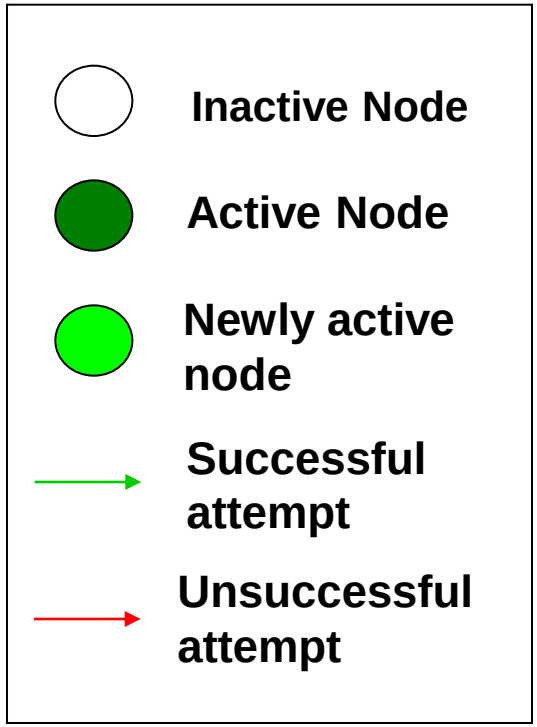
Example(IC model)



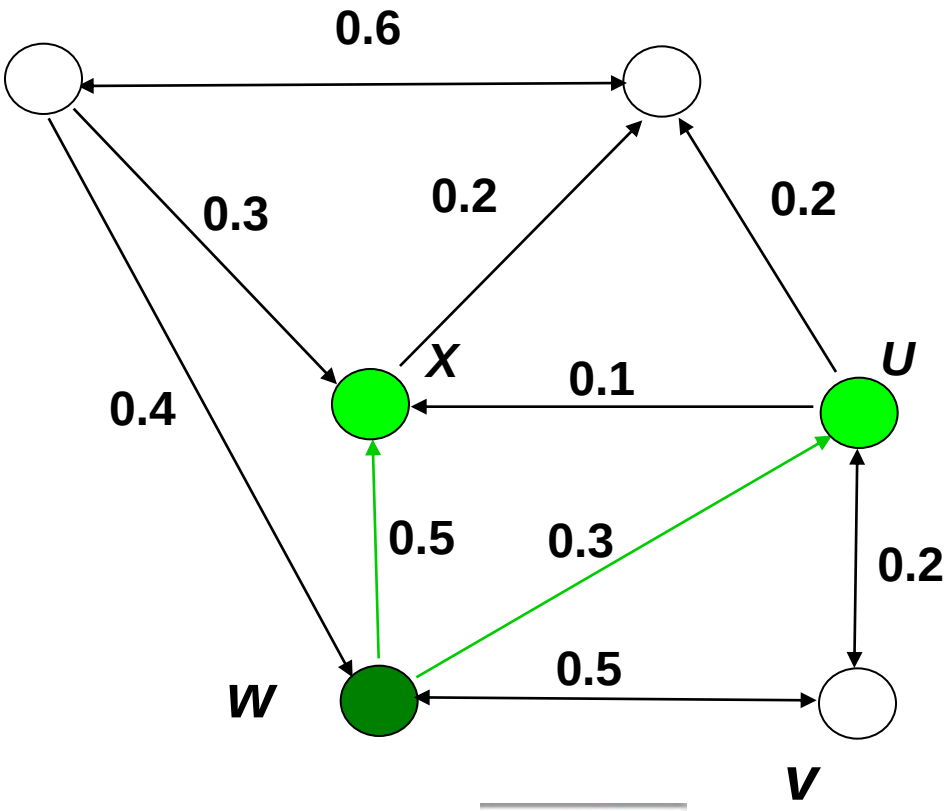


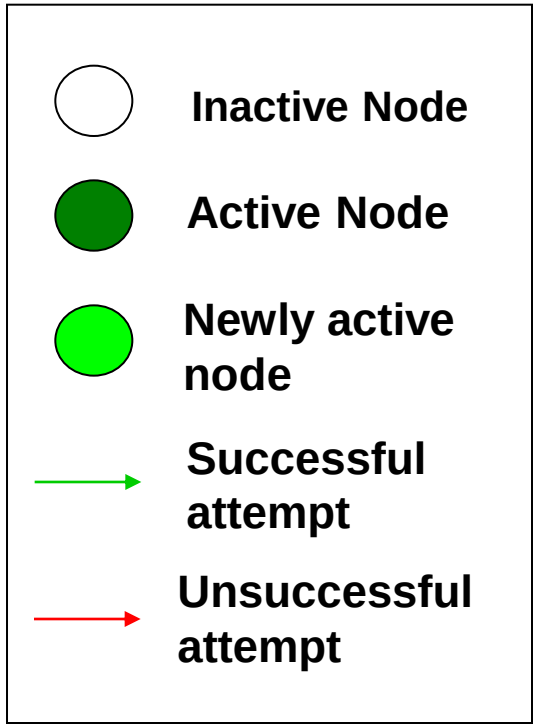
Example(IC model)



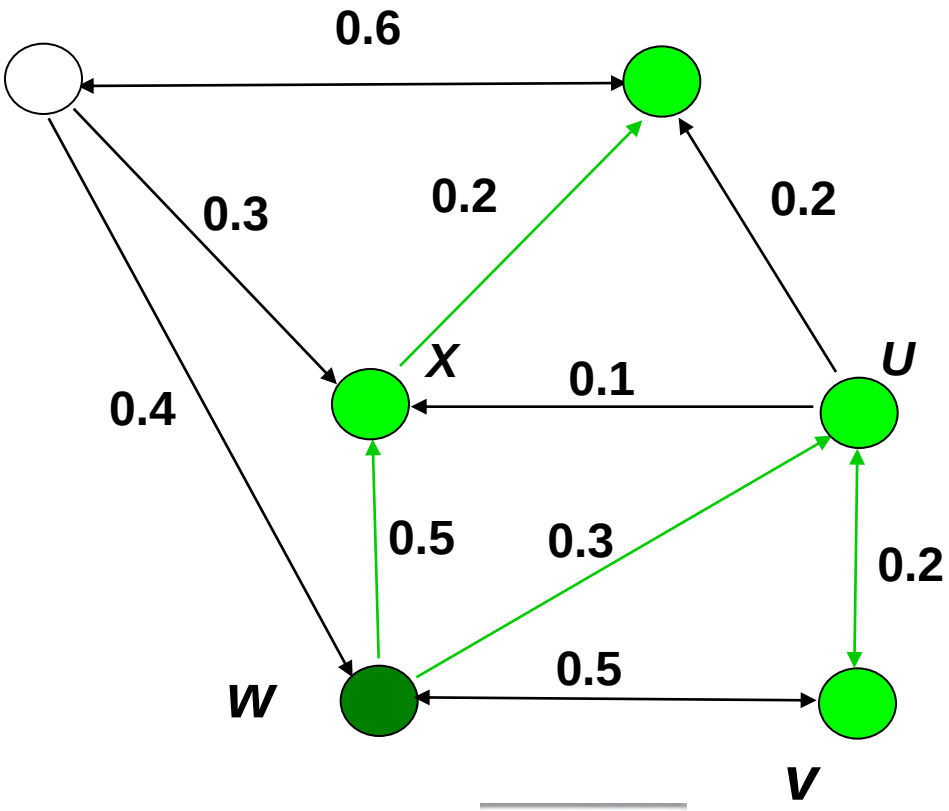


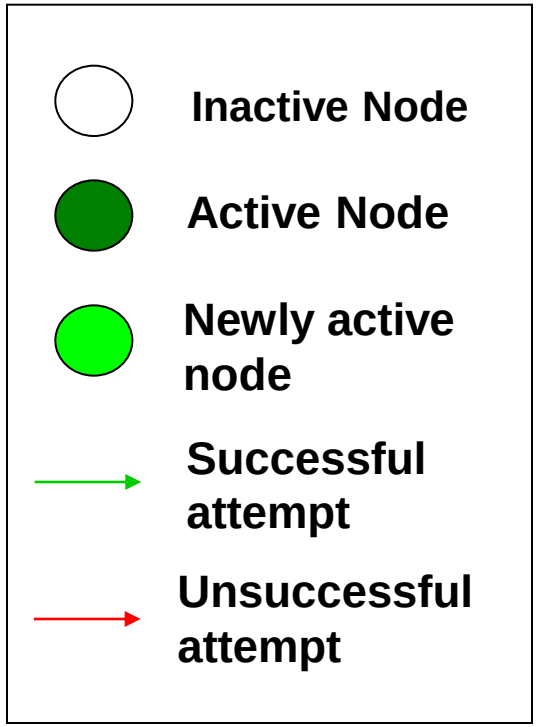
Example(IC model)



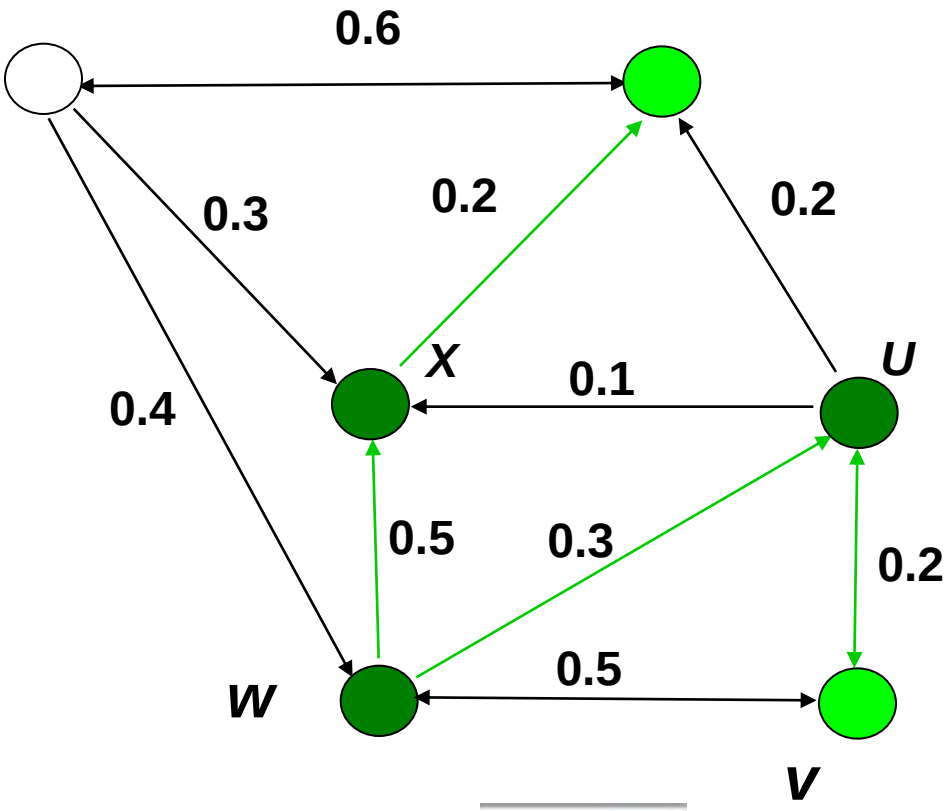


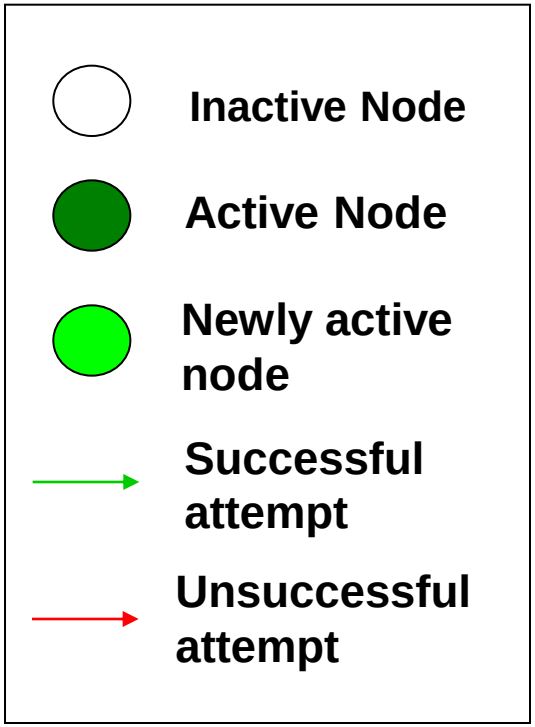
Example(IC model)



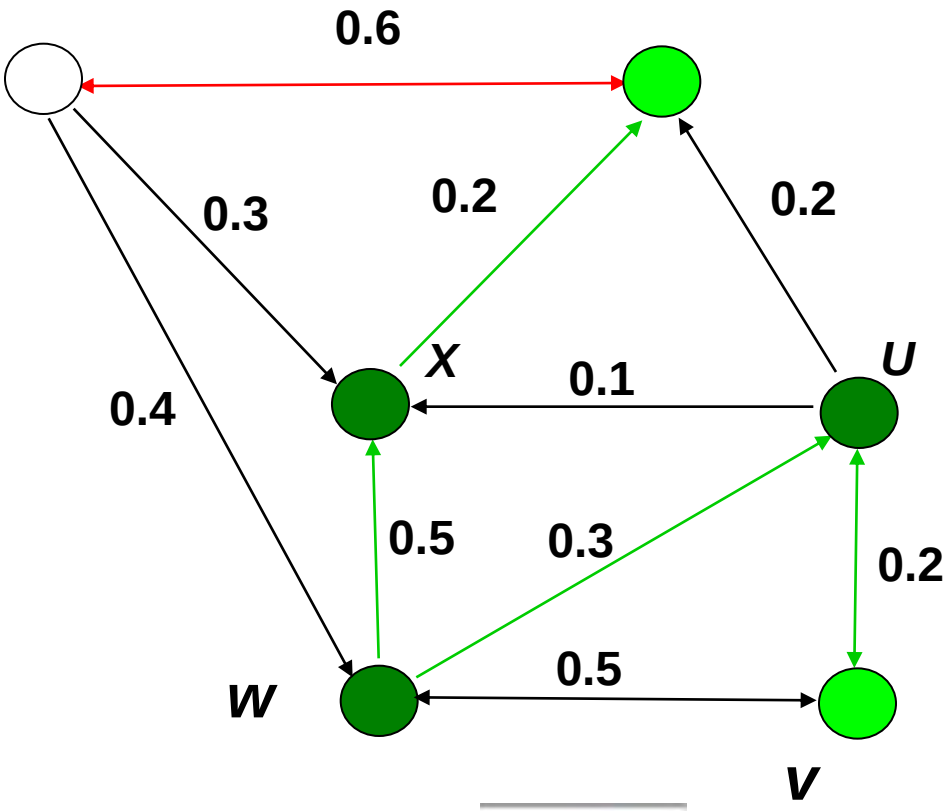


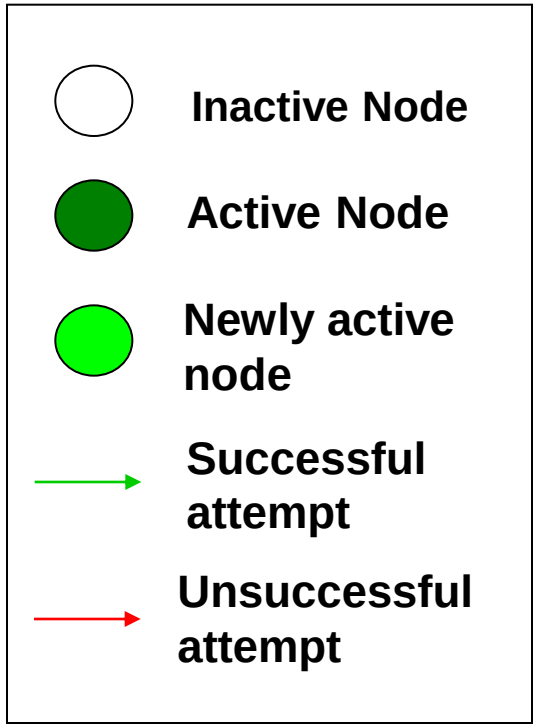
Example(IC model)



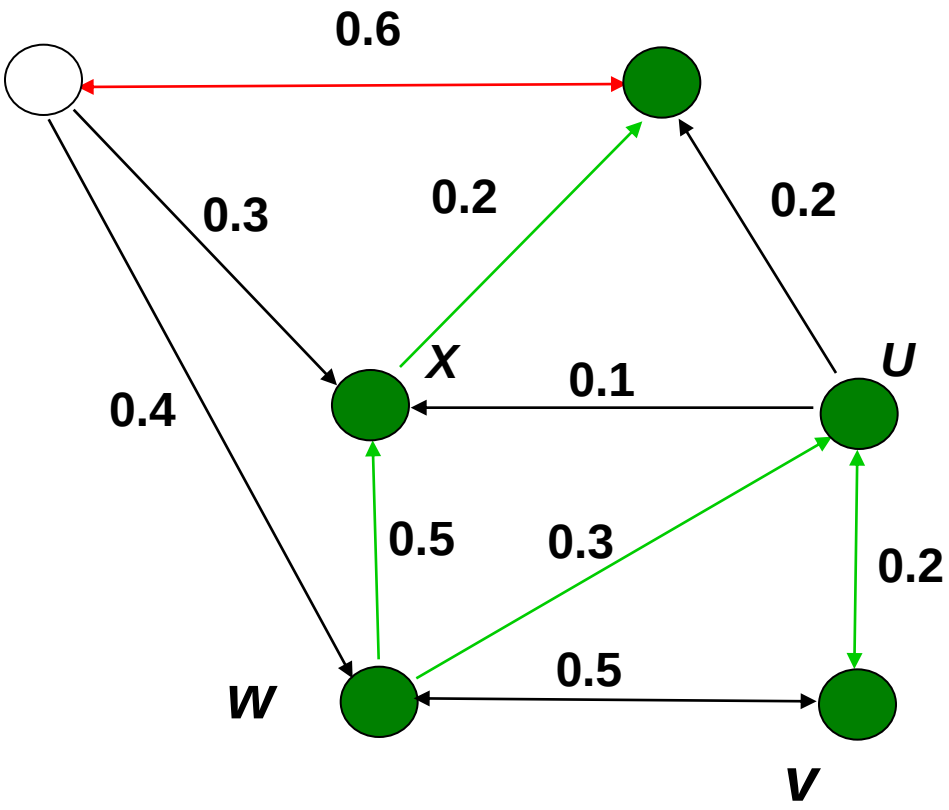


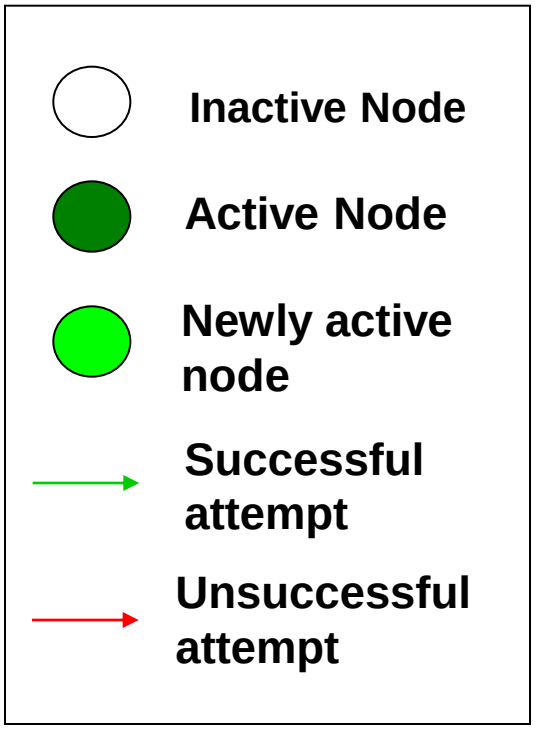
Example(IC model)





Example(IC model)





Stop!

Influence Maximization Problem

- Influence of a node set S
  - g(S) : expected number of cascade size at the end of the diffusion, where S is the initial adaptors set

- Problem:
  - Given a parameter *k* (budget), find a *k*-node set *S* to maximize g(S)
  - Constrained optimization problem with g(S) as the objective function

Properties of g(S)

- Non-negative
- Monotone: $g(S+v) \geq g(S)$
- Submodular:
  - Let *N* be a finite set
  - A set function $f: 2^N \mapsto \Re$ is submodular *iff*

$$\forall S \subset T \subset N, \forall v \in N \setminus T,$$

$$g(S+v) - g(S) \geq g(T+v) - g(T)$$

(diminishing influence)

Hardness

- For a submodular function *f*, if *f* takes non-negative value, and is monotone, finding a *k*-element set *S* for which *g(S)* is maximized is an NP-hard optimization problem [GFN77, NWF78].

- It is NP-hard to determine the optimum for influence maximization for IC model.

Approximation

- Sequential greedy Algorithm
  - Start with an empty set S
  - For k iterations:

    Add node v to S that maximizes g(S +v) - g(S).
- How good it is?
  - Theorem: The greedy algorithm is a (1 – 1/*e*)-approximation.
  - The resulting cascade size from S is at least (1- 1/*e*) > 63% of the number of nodes that optimum set S could activate.

Evaluating g(S)

- How to evaluate g(*S*+*v*)?
  - Monte-Carlo simulation is a simplest way, but it is not scalable.
- Many algorithms are proposed
  - Leskovec et. al. [KDD '07] CELF algorithm
  - Chen et. al. [KDD '10] P-MIA algorithm
- However, CELF runs too slow for large scale complex networks, and P-MIA requires too much memory.

Our Result

- We obtained a novel message passing algorithm to estimate g(S) which runs much faster than CELF and uses much less memory than P-MIA.
  - Based on a novel recursive formula to compute influence of each node
  - Experiments show that our algorithm achieves better/similar accuracy than P-MIA and CELP

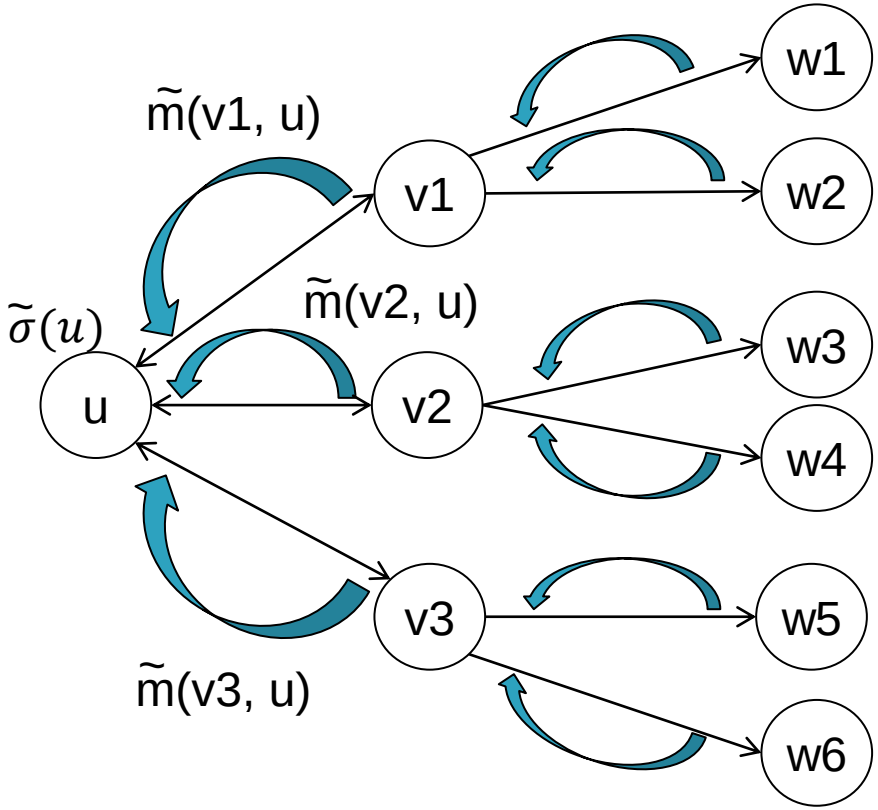
Influence Rank (IR)

- Based on this observation, we obtain Influence Rank formula

$$r(u) = 1 + \left(\sum_{v \in N^{out}(u)} P_{uv} \cdot r(v) \right).$$

- One iteration of simple IR takes $O(\sum_{v \in V} d_{out}(v))$ time.

Influence Rank (IR)



- We prove that for tree graphs, IR computes the influence $\sigma(u)$ for each node u.

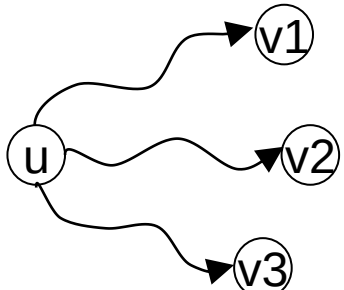
Influence Rank (IR)

- Although Simple IR runs very fast, it works only for the case that $S = \emptyset$
- Let $AP_S(u)$ be the probability that node u would be activated by the seed set S
- Then the estimates of marginal influence of u given a seed set S is

$$r(u) = (1 - AP_S(u)) \cdot \left(1 + \alpha \left(\sum_{v \in N^{out}(u)} P_{uv} \cdot r(v)\right)\right).$$

Influence Estimation (IE)

- We propose an efficient method for estimating $AP_S(v)$ given a seed set S
- Ex, let MIOA(u, S, θ) be an out-aborescence from u to other nodes consisting of paths with the highest propagation probability from u to other nodes [Chen et al. 2010]



- By generating MIOA(s_i, S, θ), $\forall\ s_i \in S$, we estimate $AP_S(v)$ according to following equation

$$AP_S(v) = \sum_{s_i \in S} AP_{s_i}(v)$$

Experiments

- Datasets
  - Six real-world datasets
  - Synthetic datasets

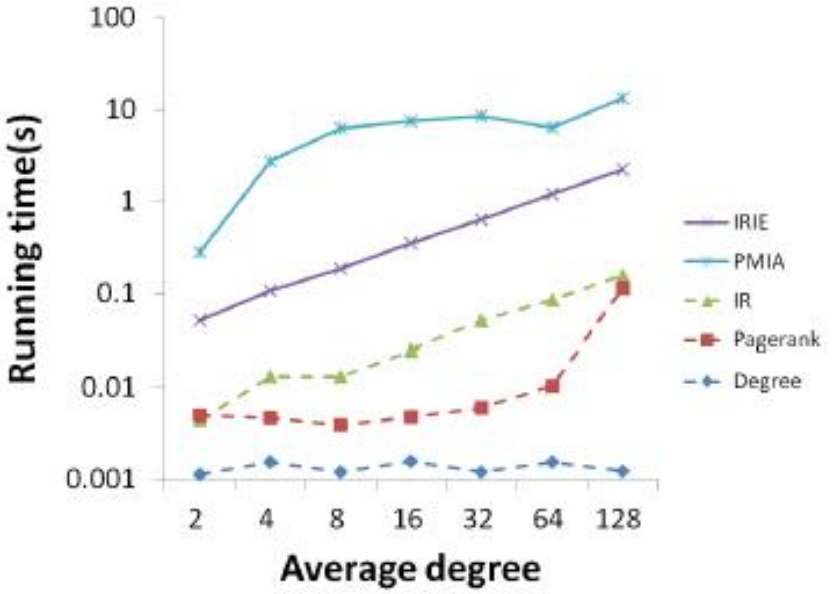
Table 1: Summary of Real-world Social Networks

Dataset	#nodes	#edges	direction
ArXiv	5K	29K	undirected
Epinions	76K	509K	directed
Slashdot	77K	905K	directed
Amazon	262K	1.2M	directed
DBLP	655K	2M	undirected
LiveJournal	4.8M	69M	directed

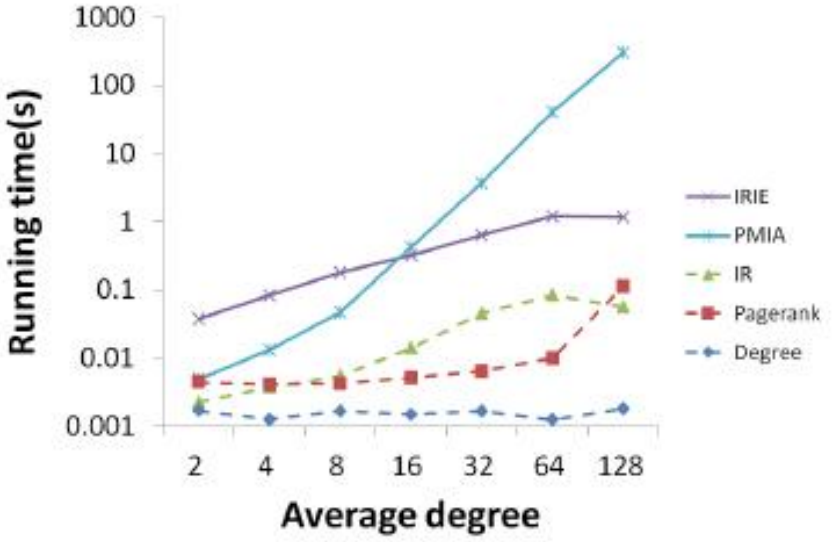
- Propagation probability models
  - Weighted cascade model (WC)
    - $P_{uv} = \frac{1}{d_v}$ where d_v is in-degree of v
  - Trivalency model (TR)
    - for each edge (u, v) , Assign P_{uv} a random probability among {0.1, 0.01, 0.001}

Experiments

- Scalability
  - Increase average degree of nodes while fixing # nodes



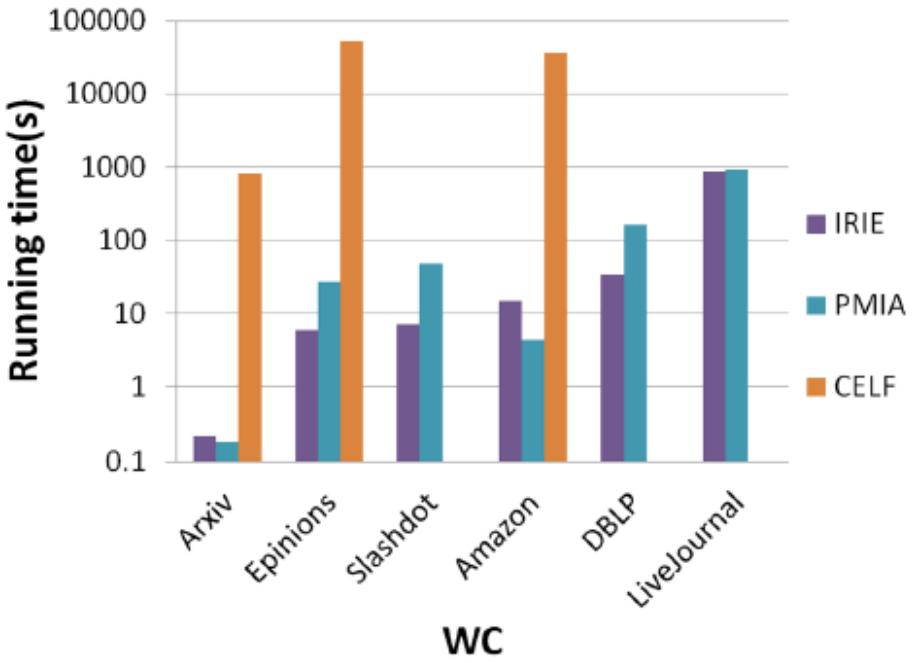
(c) Weighted Cascade



(d) Trivalency

Experiments

- Running time
  - IRIE is faster than PMIA and shows stable running time



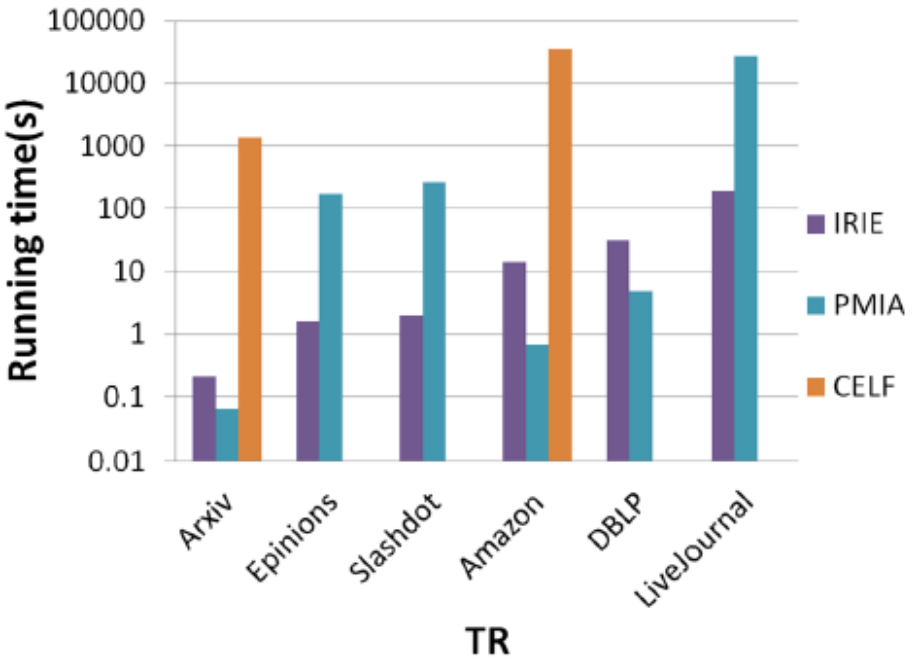


Figure 4.7: Running time of algorithms under IC model

Experiments

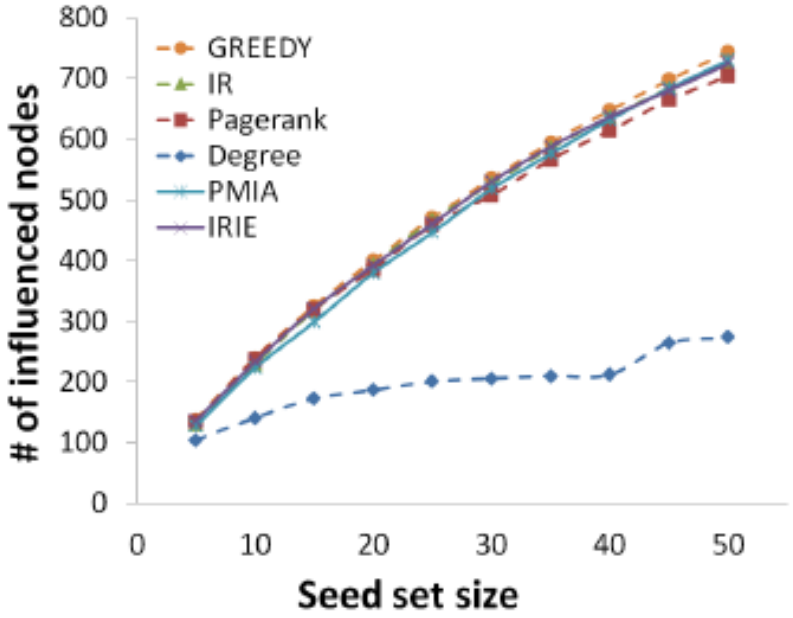
- Memory efficiency
  - IRIE is 2~7 times efficient than PMIA

Table 4: Memory usages of IRIE and PMIA

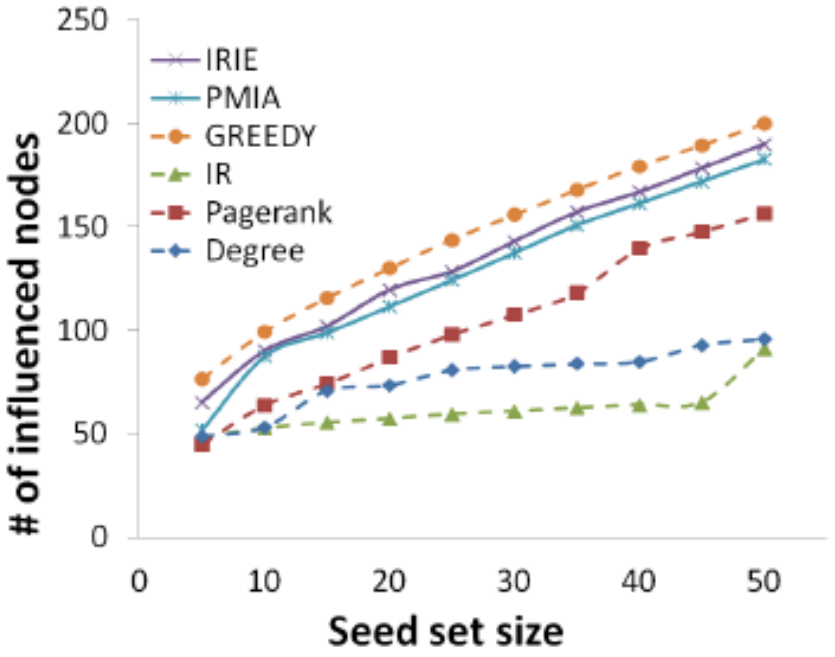
	WC			TR		
Dataset	File size	PMIA	IRIE	File size	PMIA	IRIE
ArXiv	715KB	14MB	8.7MB	582KB	10MB	8.7MB
Epinions	18MB	135MB	35MB	15MB	143MB	35MB
Slashdot	24MB	280MB	39MB	19MB	340MB	40MB
Amazon	43MB	276MB	74MB	38MB	162MB	74MB
DBLP	88MB	1.1GB	160MB	82MB	357MB	158MB
LiveJournal	2.4GB	10.1GB	3GB	2GB	16GB	3GB

Experiments

- Influence spreads



(a) Weighted Cascade



(b) Trivalency

Figure 4.2: Influence spread for ArXiv dataset

Experiments

- Influence spreads

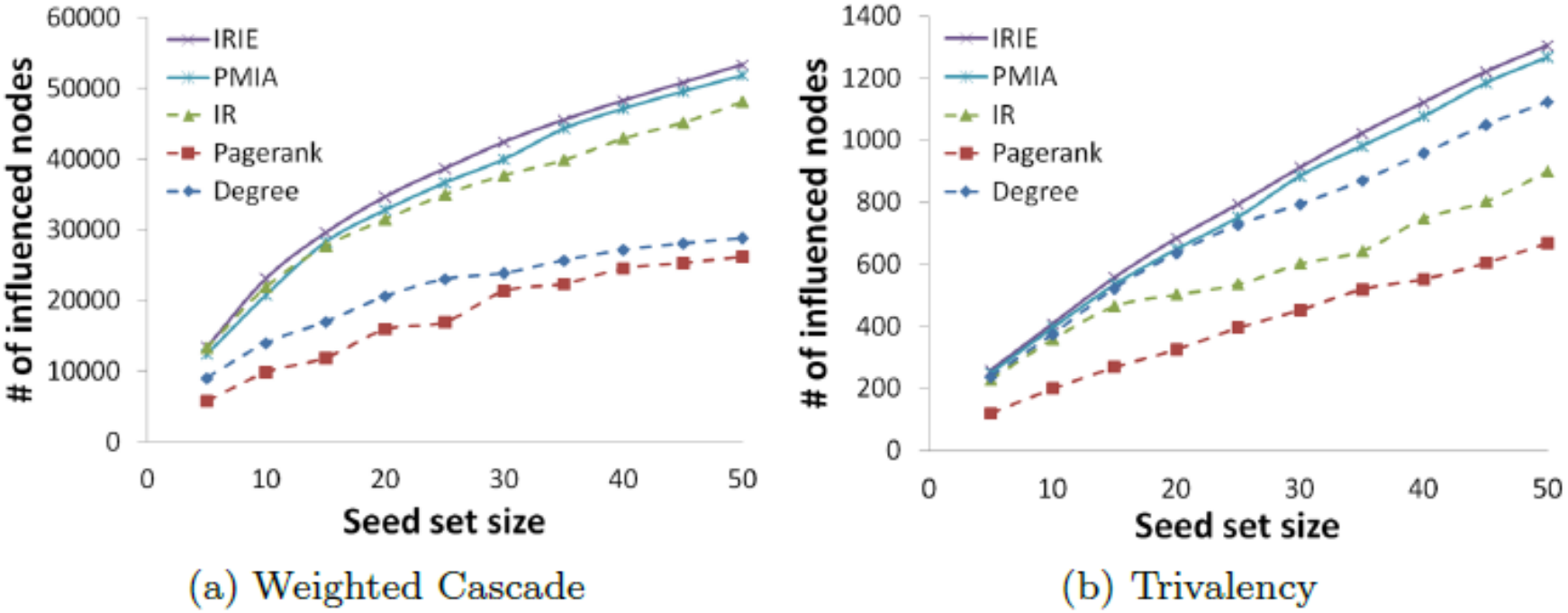


Figure 4.6: Influence spread for DBLP dataset