Enhancing Structural Diversity in Social Networks by Recommending Weak Ties

UNIVERSIDAD AUTONOMA DE MADRID

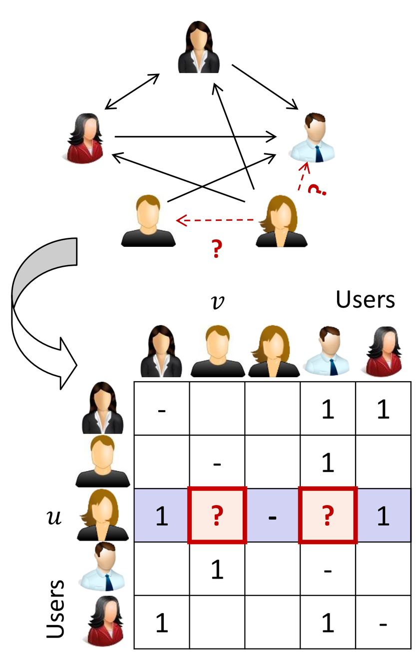
Javier Sanz-Cruzado and Pablo Castells
Universidad Autónoma de Madrid
{javier.sanz-cruzado,pablo.castells}@uam.es

IRG
IR Group @UAM

Motivation

The contact recommendation task

- Given:
  - A social network $\mathcal{G} = \langle \mathcal{U}, E \rangle$
    - $\mathcal{U}$ – Network users
    - $E \subset \mathcal{U}_*^2 = \mathcal{U}^2 \setminus \{(u,u) | u \in \mathcal{U}\}$ – Network edges
  - Neighborhoods $\Gamma(u)$ for each user $u \in \mathcal{U}$
    - $\Gamma_{in}(u) = \{v \in \mathcal{U} | (v,u) \in E\}$
    - $\Gamma_{out}(u) = \{v \in \mathcal{U} | (u,v) \in E\}$
- For each $u \in \mathcal{U}$, predict k users which might be of interest
  - $u \in U \to \hat{\Gamma}_{out}(u) = \langle u_1, u_2, \dots, u_n \rangle, u_k \in \mathcal{U} \setminus (\{u\} \cup \Gamma_{out}(u))$
- Particularities w.r.t. classic recommendation
  - Items and users are the same set
  - Users (and consequently, items) are not isolated



Accuracy at the individual level

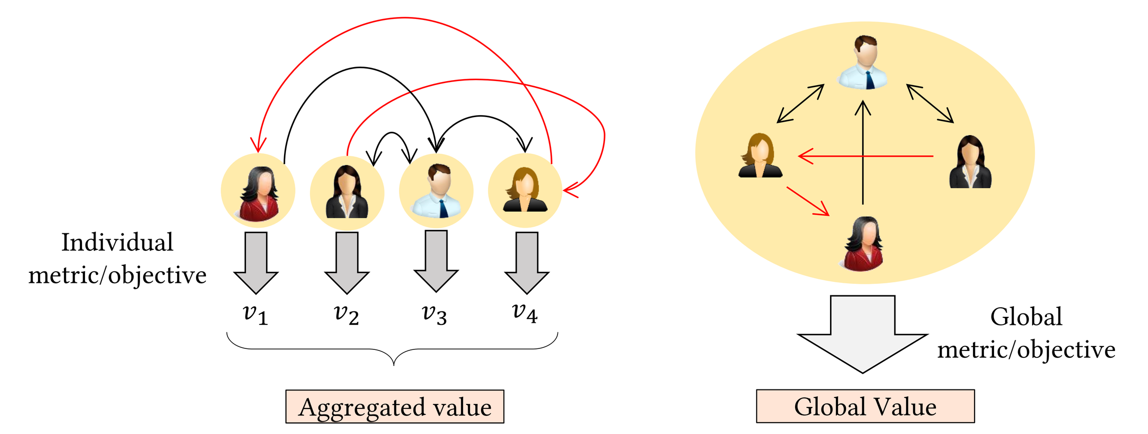
- Main focus of research and industry
- Targets the network density by correctly predicting as many edges as possible
- Measures individual gain
- However, **further qualities may enhance the value of recommendation**

Beyond the individual: global effects

- Users in networks are not isolated: few links ⟶ global effects
- Recommendations affect the shape of the network
- Opportunity to steer the evolution of the network towards desirable properties

Beyond accuracy

- Novelty & diversity
- Many notions from social network analysis
- **Structural diversity → weak ties**



Structural diversity

Weak ties

- **Strength of a tie**
  - Amount of time involved in the relationship
  - Emotional intensity
  - Intimacy (mutual confiding)
  - Reciprocal services
- Examples
  - **Strong ties:** family, close friends
  - **Weak ties:** shopkeepers, people you meet at conferences…
- Utility
  - **Strong ties:** higher reliability and availability
  - **Weak ties:** global interaction advantages, enrichment of the information flow…
- Structural notions of weak ties: non-redundant links
- Metrics applied over extended network $\mathcal{G}' = \langle \mathcal{U}, E' \rangle$
  - Assume recommendations are accepted

$$E' = E \cup \hat{E} \qquad \hat{E} = \{(u,v) \in \mathcal{U}_*^2 | u \in \mathcal{U}, v \in \hat{\Gamma}_{out}(u)\}$$

Global redundancy: Links between communities

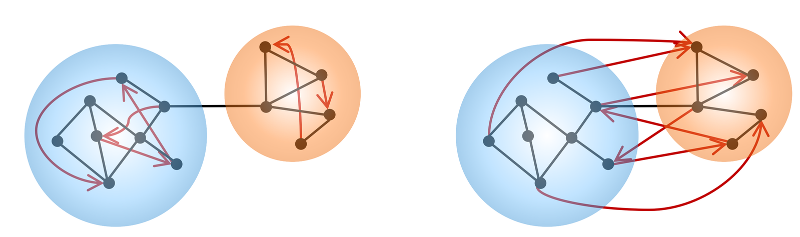
- Given a community division $\mathcal{C}$ of the network
- **Weak ties:** links between communities
- **Modularity Complement (MC)**
  - Modularity compares
    - Number of edges inside communities (strong ties)
    - Expected number of them in a random conf. graph

$$\mathrm{Mod}(\mathcal{G}'|\mathcal{C}) = \frac{\sum_{u,v \in \mathcal{U}} (A_{uv} - |\Gamma_{out}(u)||\Gamma_{in}(v)|/|E'|)\,\mathbb{1}_{[c(u)=c(v)]}}{|E'| - \sum_{u,v \in \mathcal{U}} (|\Gamma_{out}(u)||\Gamma_{in}(v)|/|E'|)\,\mathbb{1}_{[c(u)=c(v)]}}$$

  - High modularity ⟶ Few weak ties ⟶ Low structural diversity

$$\mathrm{MC}(\mathcal{G}'|\mathcal{C}) = \frac{1 - \mathrm{Mod}(\mathcal{G}'|\mathcal{C})}{2}$$

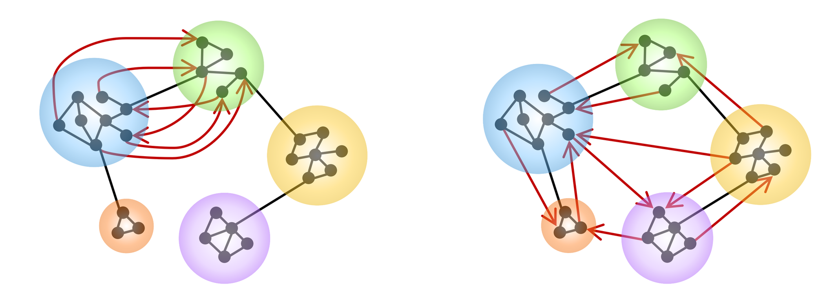
  - Limitation: it just considers the raw number of links crossing communities



a) Low MC   b) High MC

- **Community edge Gini complement (CEGC)**
  - Considers redundancy between weak ties
  - Analyzes distribution of links crossing communities
    - Low CEGC⟶ Skewed distribution ⟶ Low diversity
    - High CEGC⟶ Balanced distribution ⟶ High diversity
  - Based on the Gini Index
    - n_{ij}: Number of links between communities c_i, c_j
    - $X(\mathcal{G}'|\mathcal{C}) = \{n_{ij} | i \neq j\} \cup \{\sum_i^{|\mathcal{C}|} n_{ii}\}$,
    - $N = |X(\mathcal{G}'|\mathcal{C})| = (|\mathcal{C}| - 1)|\mathcal{C}| + 1$
    - Sorted set: $x_1 \leq x_2 \leq \cdots \leq x_N$

$$\mathrm{CEGC}(\mathcal{G}'|\mathcal{C}) = 1 - \frac{1}{N-1} \sum_{j=1}^{N} (2j - N - 1) \frac{x_i}{|E'|}$$



a) High MC, low CEGC   b) High MC, high CEGC

Local redundancy: Transitive closure

- **Triadic closure:** smallest unit of structural redundancy
- **Clustering coefficient complement (CCC)**

a) Open triad   b) Closed triad

$$\mathrm{CC}(\mathcal{G}') = \frac{|\{(u,v,w) | (u,v),(v,w),(u,w) \in E'\}|}{|\{(u,v,w) | (u,v),(v,,w) \in E' \wedge u \neq v\}|}$$

$$\mathrm{CCC}(\mathcal{G}') = 1 - \frac{|\{(u,v,w) | (u,v),(v,w),(u,w) \in E'\}|}{|\{(u,v,w) | (u,v),(v,,w) \in E' \wedge u \neq v\}|}$$

Recommended edges ($\hat{E}$)
Original edges (E)

Recommendation experiments

Data

- 2 Twitter samples (directed networks)
- Interaction graphs: $(u,v) \in E \iff u$ mentions, retweets v
- Temporal split
- Community detection algorithm: Louvain

	Complete network		Training network			Test network	
Dataset	#Users	#Edges	#Users	#Edges	#Comm.	#Users	#Edges
1 month	10,019	234,869	9,528	170,425	8	7,902	57,846
200 tweets	10,000	164,653	9,985	137,850	10	5,652	21,598

Algorithms

- **Neighborhood based:** Most Common Neighbors, Adamic-Adar, Jaccard
- **Random walks:** Personalized SALSA
- **Content-based:** Centroid CB
- **Classic recommendation:** Implicit Matrix Factorization (MF)
- **Baselines:** random, popularity

	Recommender	Optimal parameters	P@10	R@10	MC	CEGC	CCC
1 month	Implicit MF	$k = 260, \lambda = 150, \alpha = 40$	**0.0625**	**0.1060**	0.1550	0.0447	0.9766
	Personalized SALSA	Authorities, $\alpha = 0.99$	0.0577	0.0990	0.1656	0.0447	0.9819
	Adamic-Adar	und, in, und	0.0505	0.0697	0.1487	0.0413	0.9748
	MCN	und, in	0.0476	0.0647	0.1461	0.0403	0.9746
	Popularity	-	0.0234	0.0409	**0.2947**	0.0613	**0.9890**
	Jaccard	und, in	0.0169	0.0209	0.1464	0.0434	0.9652
	Centroid CB	in	0.0156	0.0198	0.1652	0.0498	0.9627
	Random	-	0.0006	0.0009	0.2797	**0.0901**	0.9839
	Training graph	-	-	-	*0.1464*	*0.039*	*0.9829*
200 tweets	Implicit MF	$k = 300, \lambda = 150, \alpha = 40$	**0.0236**	**0.0589**	0.2132	0.1326	0.9520
	Adamic-Adar	und, in, und	0.0233	0.0540	0.2076	0.1180	0.9447
	MCN	und, in	0.0222	0.0499	0.2048	0.1138	0.9433
	Personalized SALSA	Authorities, $\alpha = 0.99$	0.0208	0.0516	0.2369	0.1412	0.9594
	Centroid CB	in	0.0157	0.0333	0.2154	0.1251	0.9182
	Jaccard	und, in	0.0132	0.0306	0.2041	0.1195	0.9065
	Popularity	-	0.0098	0.0221	**0.3371**	0.1559	**0.9822**
	Random	-	0.0003	0.0007	0.3317	**0.2276**	0.9795
	Training graph	-	-	-	*0.2081*	*0.1134*	*0.9559*

- **How do state of the art algorithms perform in terms of structural diversity?**

Effect on information diffusion

Hypothesis

The more structurally diverse is the recommendation, the more diverse and novel (non-redundant) will be the information flow through the network

Experiment description

- Start with a well-behaved baseline → Implicit MF (most accurate method)
- Rerank baseline to enhance a structural metric of the network
- Simulate the flow of information through the extended network $\mathcal{G}'$
- Analyze properties of diffusion (speed, novelty & diversity)

Data

- Same networks as the ones used for the recommendation experiments
- **Information to propagate:** Tweets
  - originally published after the temporal split
  - containing hashtags which appear in (at least) 25 different tweets (avoid noise)

#Dataset	#Tweets	#Hashtags (unique)
1 month	87,837	110,578 (1115)
200 tweets	21,513	24,623 (378)

Metrics enhancement

- Enhance a global property μ of the network
- Rerank baseline recommendation by greedy maximization of objective function

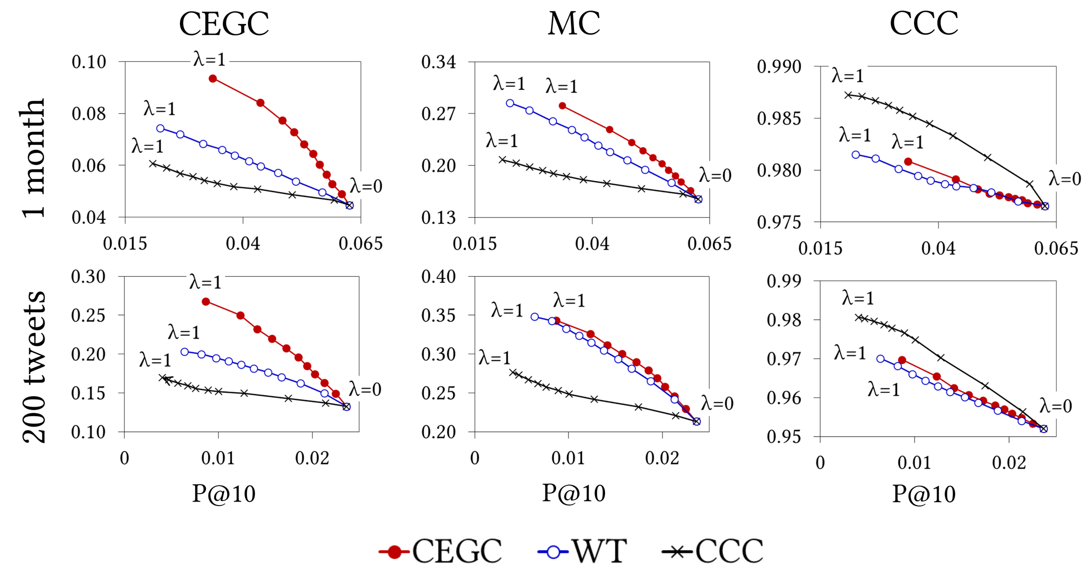
$$\phi(S, f, \mu, \lambda) = (1 - \lambda) \sum_{u \in \mathcal{U}} \sum_{(u,v) \in S_u} f(u,v) + \lambda\, \mu(\mathcal{G}'_S)$$

- **Algorithm:** Global greedy reranking

```
Input:  Ê ⊂ 𝒰*²      original recommendations
        f: Ê → ℝ     original recommendation ranking function
        μ            metric to optimize
        k            diversification cutoff
        λ ∈ [0,1]    degree of diversification
        𝒢 = ⟨𝒰, E⟩   training graph
Output: S            modified recommendations (a set of ordered lists)
begin
 S ← sort(Ê, f) // Edges are grouped by source node and sorted by f
 for u ∈ 𝒰 do
  for i ← 1 to k do
   j₀ ← arg max_{j:k<j≤|S_u|} φ(j|S, u, i, f, μ, λ) // S_u ≡ ranking for user u in S
   if φ(j₀|S, u, i, f, μ, λ) > φ(i|S, u, i, f, μ, λ) then swap(S_u, i, j₀)
 return S
end

Function φ(j|S, u, i, f, μ, λ)  // The dual objective function
begin
 return (1 − λ) norm(f(S_u[j])) + λ norm (μ (𝒢′_{S⟨u:i/j⟩@k}))
end
```

- Metrics for the different structural diversity rerankers (Implicit MF)



Information diffusion properties

- Notation
  - $\mathcal{H}$: Set of all hashtags
  - A tweet i is defined as a subset of $\mathcal{H}$
  - At time t, u has received the tweets $\mathcal{M}_u(t)$, containing the hashtags $\mathcal{H}_u(t)$
  - At time t, u has published $\mathcal{M}_u^0(t)$, , containing the hashtags $\mathcal{H}_u^0(t)$
- **Speed**
  - Most analyzed network efficiency feature in diffusion processes
  - How many tweets are propagated and received?

$$\mathrm{speed}(t) = \sum_{u \in \mathcal{U}} |\mathcal{M}_u(t)|$$

Novelty and diversity

- Measured in terms of hashtags
- **Novelty**
  - How new is the information received by users?
  - External hashtag rate (EHR)

$$\mathrm{EHR}(t) = \frac{\sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{M}_u(t)} |i \setminus \mathcal{H}_u^0(t)|}{\sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{M}_u(t)} |i|}$$

- **Diversity**
  - Are hashtags evenly distributed over the population?
  - Potential for diminishing filter bubbles
  - Hashtag Gini complement (HGC)

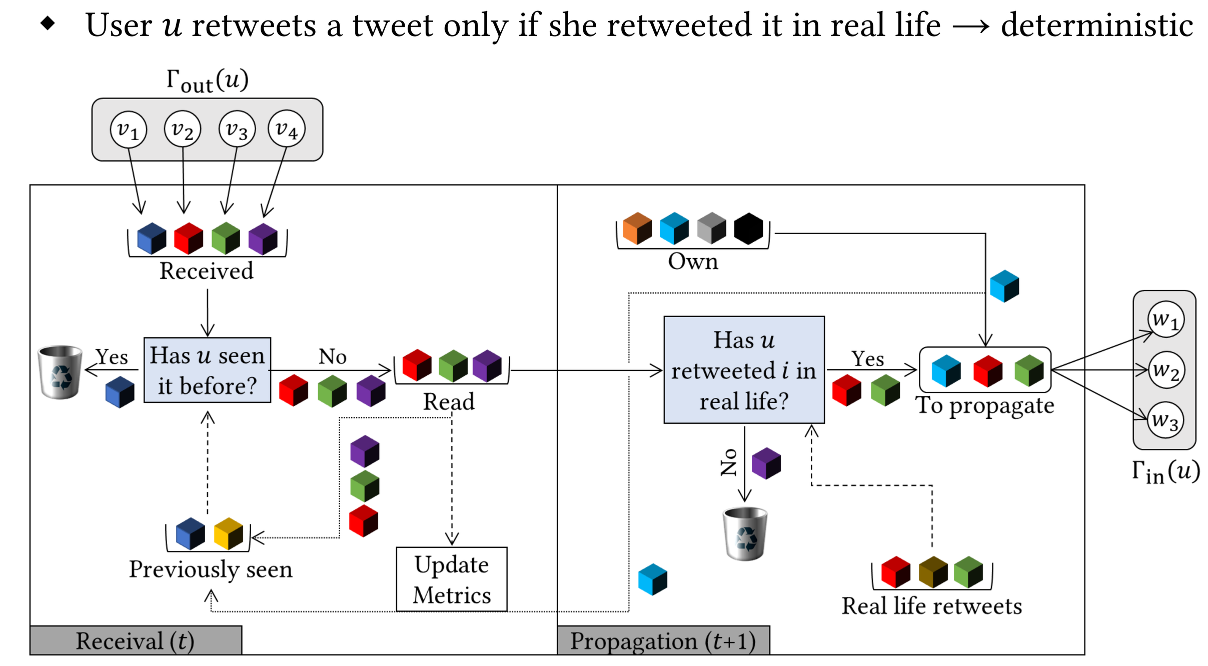
$$\mathrm{HGC}(t) = 1 - \frac{1}{|\mathcal{H}| - 1} \sum_{j=1}^{|\mathcal{H}|} (2j - |\mathcal{H}| - 1)\, p(h_j | t)$$

$$p(h_j | t) = \frac{|\{u \in \mathcal{U} | h \in \mathcal{H}_u(t)\}|}{\sum_{h^* \in \mathcal{H}} |\{u \in \mathcal{U} | h^* \in \mathcal{H}_u(t)\}|}$$

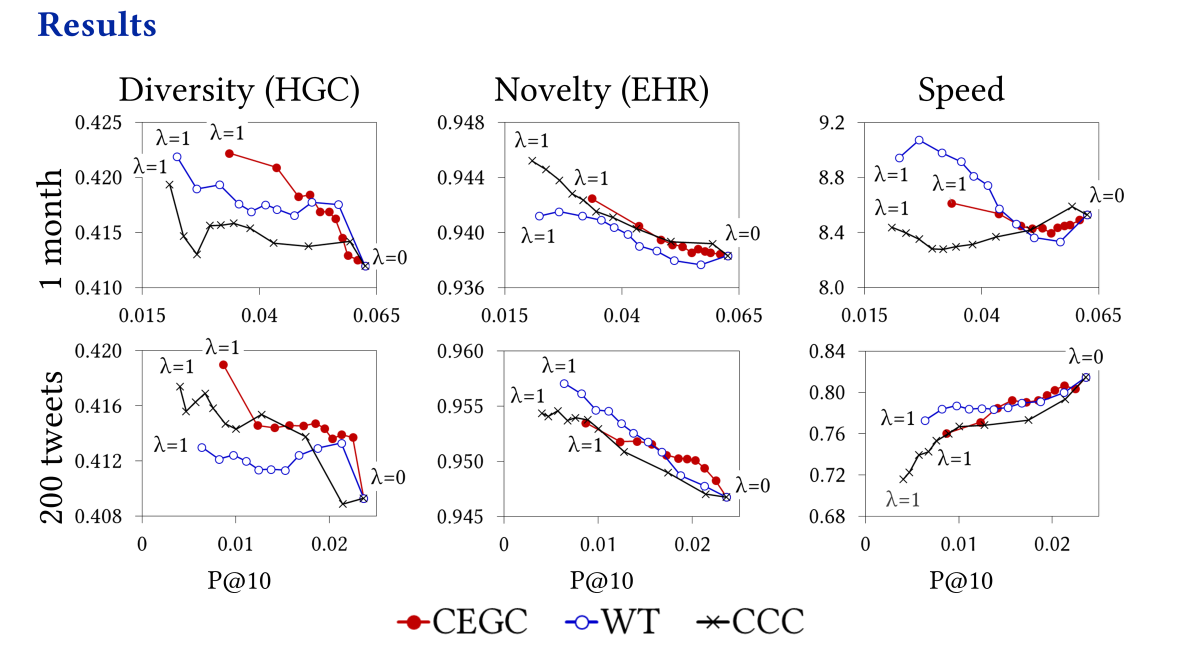
where $p(h_1|t) \leq p(h_2|t) \leq \cdots \leq p(h_{|\mathcal{H}|}|t)$

Protocol

- Information is propagated to all followers
- User u retweets a tweet only if she retweeted it in real life ⟶ deterministic



Results



Conclusions

- Information diversity is improved by enhancing structural diversity properties of the network
  - Potential relevance in mitigating filter bubbles
- CEGC provides the best trade-off between accuracy, structural properties and information diversity
- Recommending weak ties improves the novelty of the information received by the different users

12th ACM Conference on Recommender Systems (RecSys 2018)