Hot set identification for Social network applications

Michele Colajanni
Claudia Canali
Riccardo Lancellotti

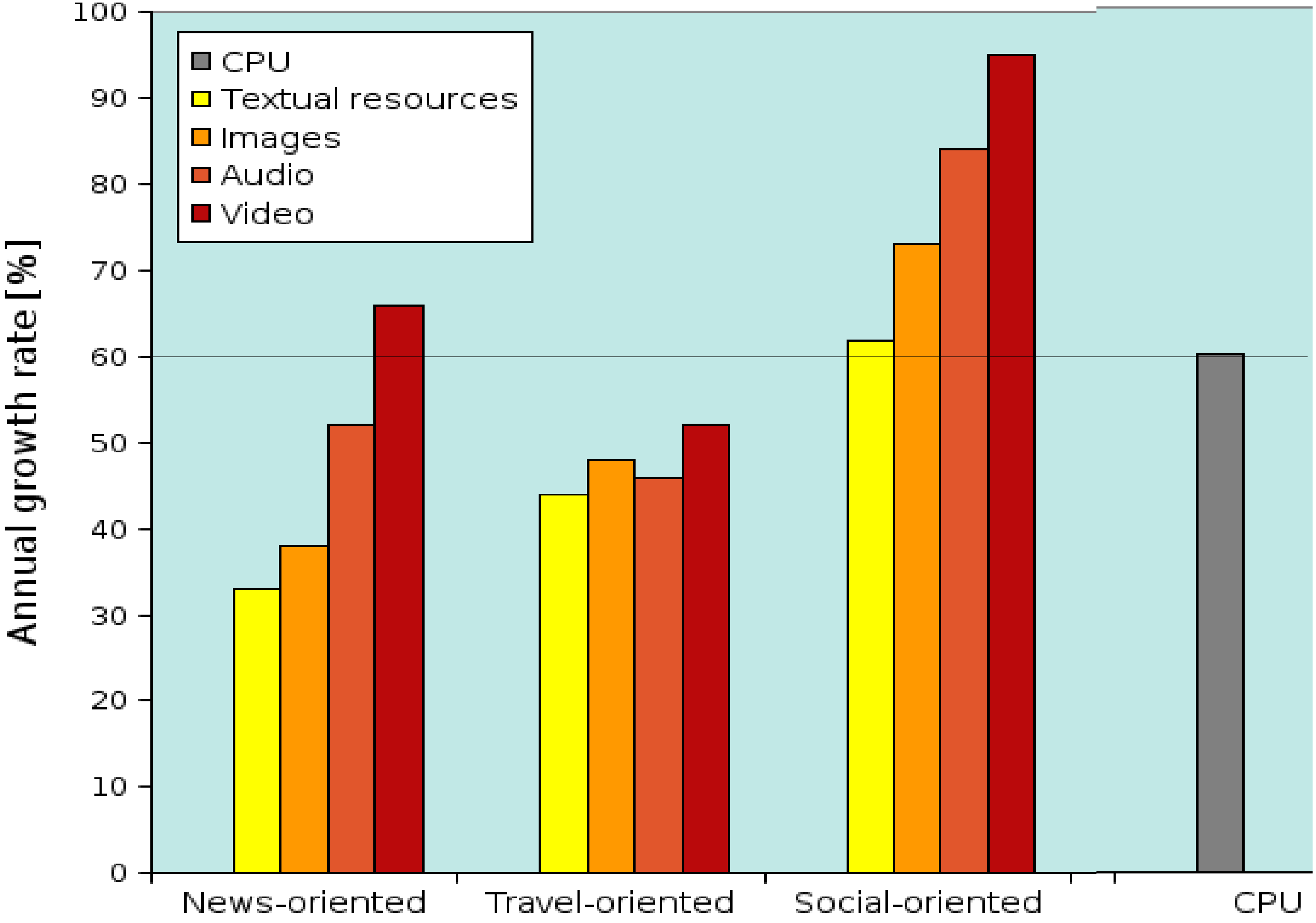
University of Modena and Reggio Emilia

Future Web Scenarios

- **Community-based services**
  - Social networking: support for user interaction be the killer of future Web
  - Rich-media content
  - Presence of Mobile User access

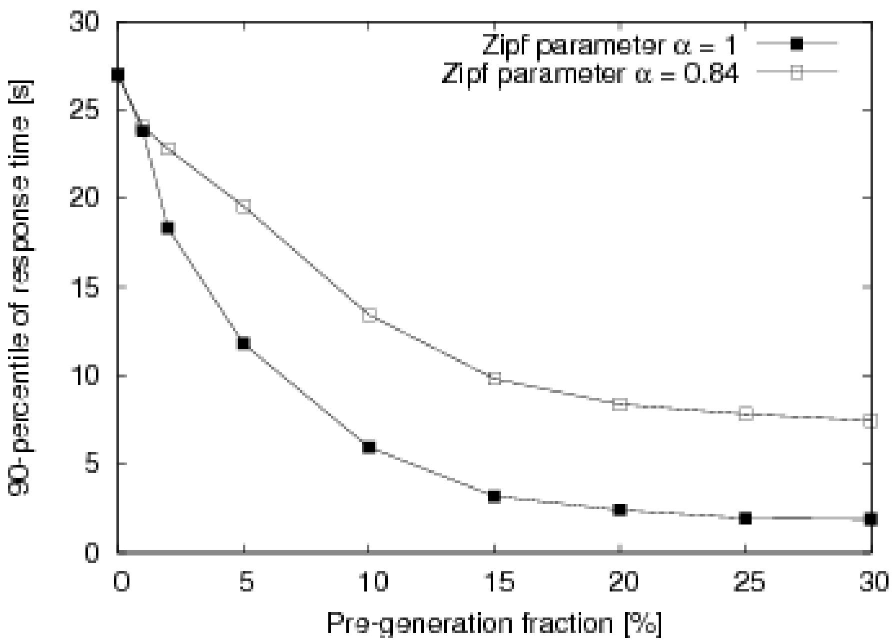
- **Workload evolution in the next five years**
  - Computational demand will grow faster than CPU power (Moore's Law)

Expected growth of computational demands



Motivations for content management

- **Content management**
  - Content replication
  - Caching
  - CDN delivery
  - Resource pre-generation



- **→ Need to identify the Hot set of popular resources**
  - Variability in workload characteristics
  - Rapid variations in access patterns
  - Workload dynamics related to social interactions
- **→ Need for algorithms providing early and fast detection of popular resources.**
- **→ Stable performance are not an optional**

Proposal: Algorithms for Hot set identification

- **The algorithm must identify the set HS(t)**
  - Hot set is evaluated periodically with interval Δt
  - HS(t) will receive the highest number of accesses in the interval [t, t+Δt]
  - HS(t) subset of R(t), working set at time t

- **An algorithm must:**
  - Estimate $p_r(t)$, where $p_r(t)$ is the popularity of resource r in interval [t, t+Δt]
  - Sort R(t) according to $p_r(t)$

- **→ HS(t) is the top fraction of sorted set R(t)**

Proposed algorithms

- **Critical task for every algorithm**
  - Evaluation of $p_r(t)$
- **Three classes of innovative algorithms**
  - Predictive
  - Social-aware
  - Predictive-Social
- **Comparison with existing solutions**

Existing algorithms

- **Focus on the time interval [t-Δt, t]**
  - $d_r(t)$ is the number of access to resource r in interval [t-Δt, t]

- **Access frequency as a measure of resource popularity**
  - $p_r(t)=d_r(t)/\Delta t$

- **Similar to frequency-based algorithms already used for cache replacement**

Predictive algorithms

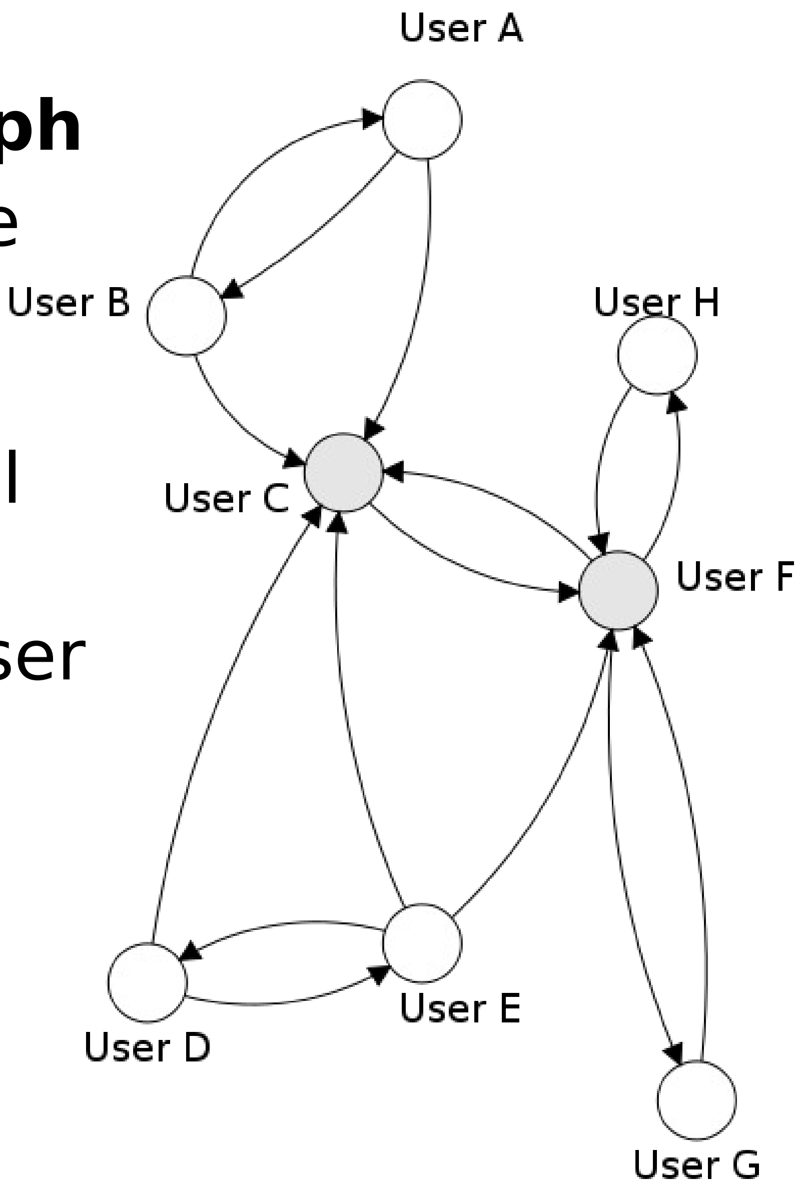
- **History of past accesses to resource r represented as a time series:**
  - $D_r(t)=\{d_r(t), d_r(t-\Delta t), \ldots, d_r(t-(n-1)\Delta t)\}$
  - $d_r(t)$ is number of accesses to resource r in interval [t-Δt, t], $d_r(t-\Delta t)$ refer to [t-2Δt, t-Δt], ...

- **Use of an EWMA model for prediction:**
  - $d_r^*(t,t+\Delta t)=\gamma d_r^*(t,t+\Delta t)+(1-\gamma)d_r(t)$
  - $\gamma=2/n$, where n is the time series length

- **Other prediction models are possible**

Social-aware algorithms

- **Social network can be represented as a directed graph**
  - Reverse contact represent the popularity of a user within the social network
  - User navigation exploits social links
  - Strong correlation between user popularity and popularity of uploaded resources
  - → Popular users are likely to publish popular content



Social-aware algorithms

- **Popularity estimation based on user reverse contacts**
  - $c_r(t)$ connection degree of user that uploaded resource r
  - $c_{max}(t)$ maximum connection degree

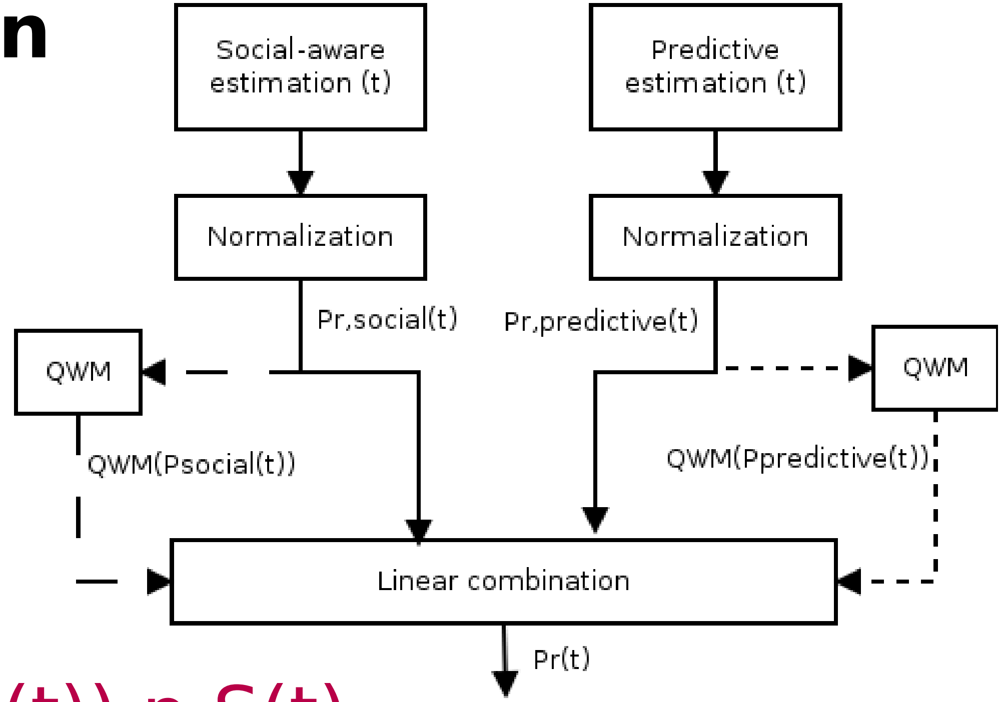
- **The model includes also the effect of resource aging**
  - $a_r(t)$ age of resource r (time since resource upload)
  - $p_r(t)=c_r(t)/(c_{max}(t)\ a_r(t))$

Predictive-Social algorithms

- **Most innovative class of algorithms**
  - Merges information from two sources:
  - Prediction
  - Social information
- **Need for a reliable way to merge two completely different sets of data**
  - Different value ranges
  - Different probability distributions
- **Use of a robust weighting function**
  - Two-sided quartile weighted median
  - Given distribution P(t):
  - $QWM(P(t))=(Q_{25}(P(t))+2Q_{50}(P(t))+Q_{75}(P(t)))/4$

Predictive-Social algorithms

- **Merging social-aware and predictive information**
  - $p_r P(t)$ → predictive
  - $p_r S(t)$ → social
  - $\delta(t)$ → weight



- **That is:**
  - $p_r(t)=\delta(t)\, p_r P(t) + (1-\delta(t))\, p_r S(t)$
  - $\delta(t)=QWM(PS(t))/(QWM(PS(t)) + QWM(PP(t)))$

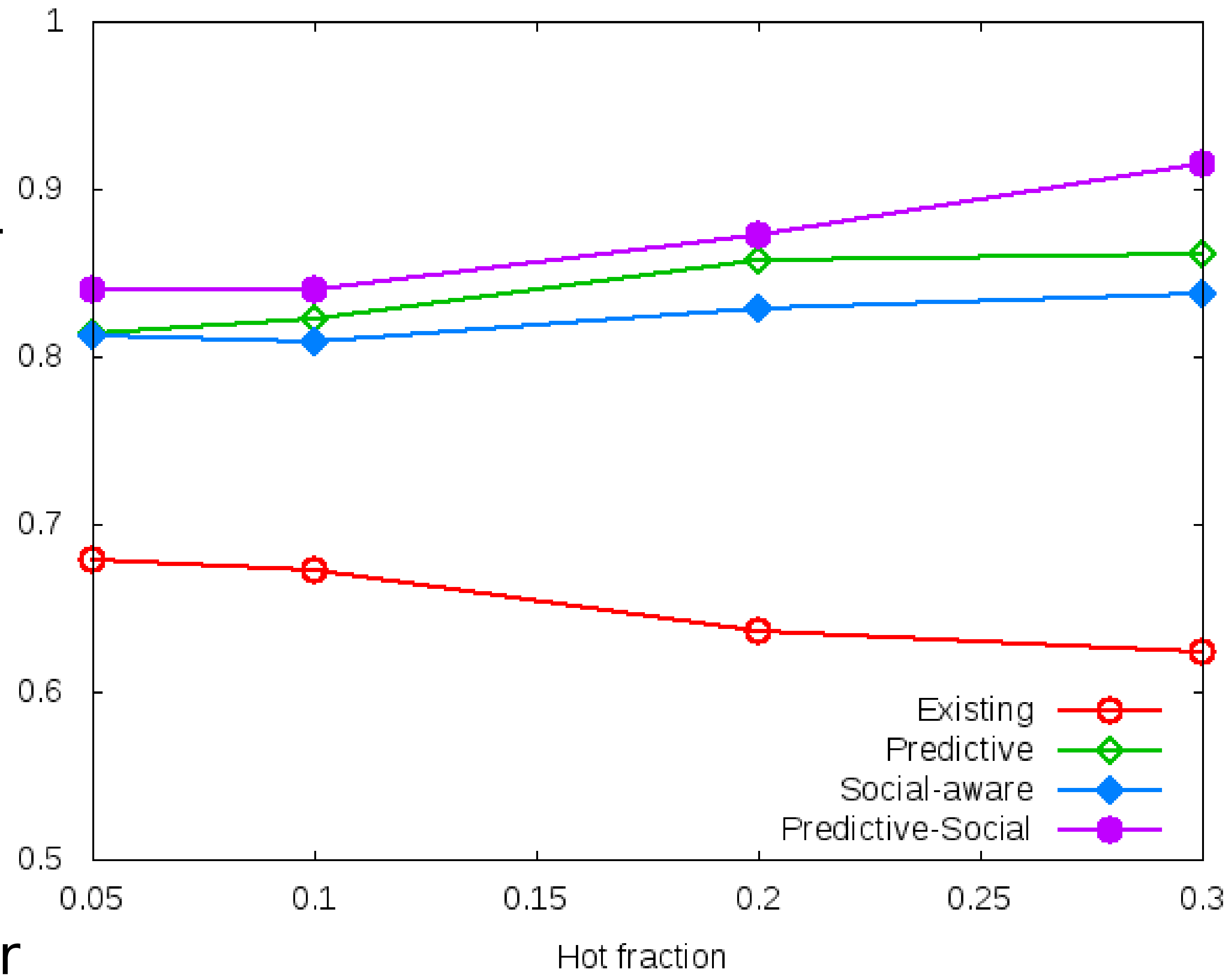
Experimental setup

- **Simulation based on Omnet++ framework**
  - User population up to 20000 units
  - Average of 100 requests/sec
  - 12 hours of simulated time
  - Δt=20minutes
  - Main metric: accuracy=$|HS(t) \cap HS^*(t)|/|HS^*(t)|$

Parameter	Range	Default
Hot fraction [%]	5%-30%	20%
Upload percentage [%]	1%-20%	5%
User/resource popularity correlation	0.6-0.8	0.7

Performance evaluation

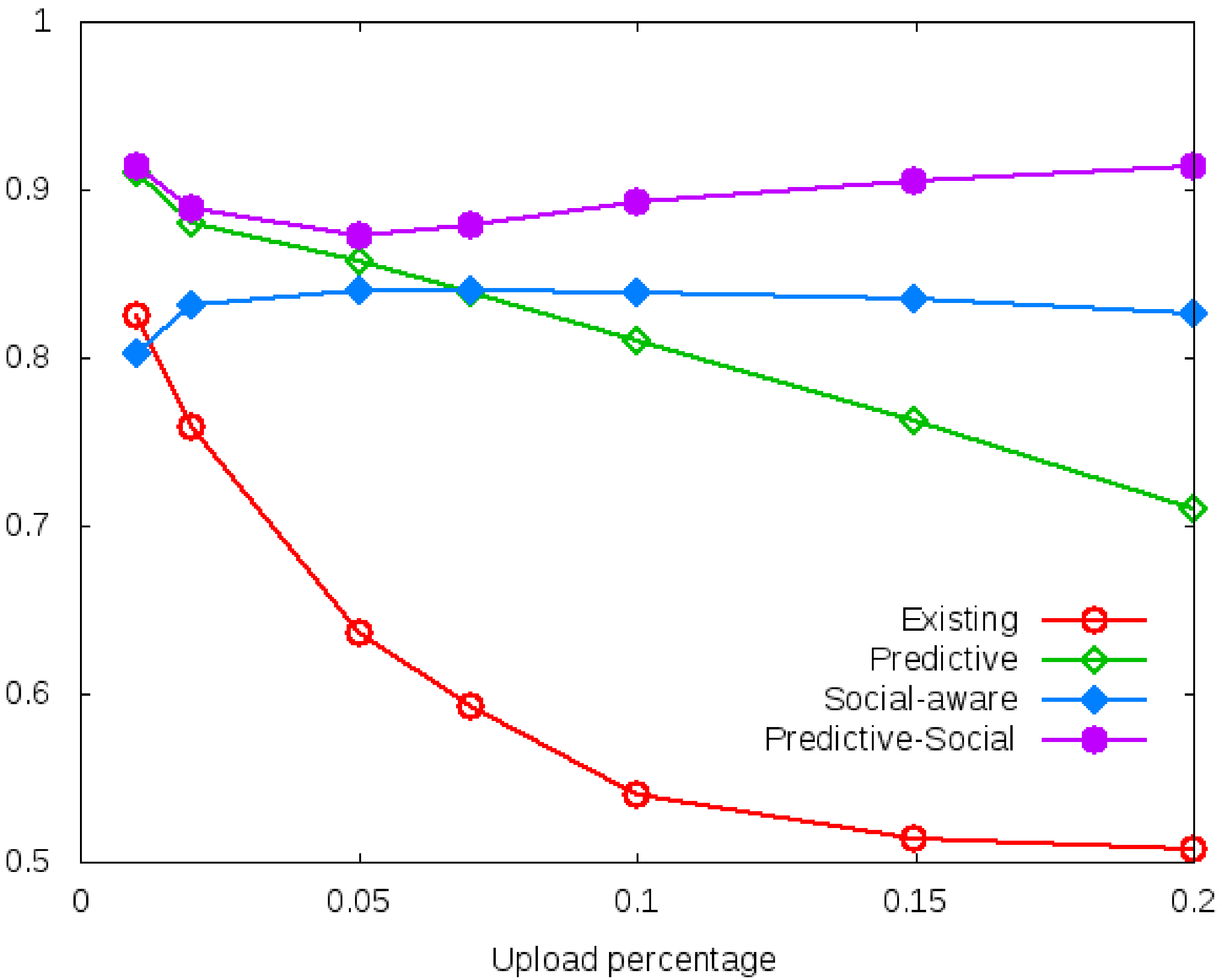
- Existing algorithms can be improved
- Predictive and social-aware algorithms provide significant improvement
- Merging prediction and social information provides further benefits
- Results are similar for every considered hot set size



→ Need to evaluate performance stability

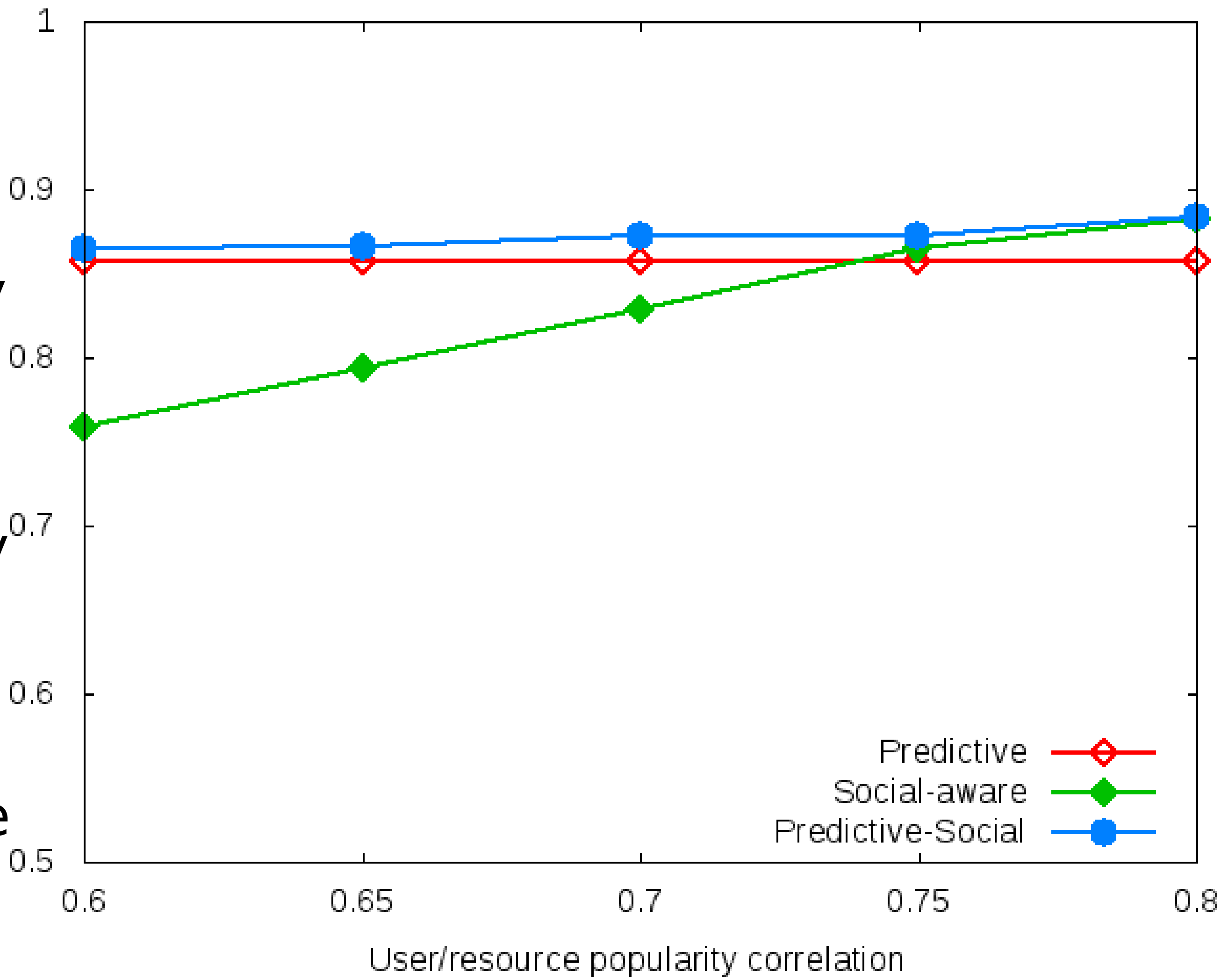
Sensitivity to workload dynamics

- Existing algorithms cannot cope with large amount of uploads
- Prediction is highly sensitive to upload percentage
- Social-aware algorithm is not sensitive to workload dynamics
- Predictive-Social algorithm provides stable performance



Sensitivity to social parameters

- Prediction is not affected by social phenomena
- Social-aware is highly sensitive to the correlation between user and resource popularity
- Predictive-Social algorithm provides stable performance



Conclusions

- **Content management will be fundamental for future social network applications**
  - Need to identify the Hot set
  - Must cope with novel challenges (social interaction, short resource lifespan, ...)
- **Need for high accuracy and stable performance**
- **Three classes of algorithms**
  - Predictive → sensitive to workload dynamics
  - Social-aware → sensitive to social dynamics
  - Predictive-Social → stable results
- **Future work**
  - Experiments with real social network traces *(any help is appreciated)*

Hot set identification for Social network applications

Michele Colajanni, Claudia Canali
Riccardo Lancellotti

riccardo.lancellotti@unimore.it

University of Modena and Reggio Emilia