Content Delivery supported by social-network awareness – Part I

Irene Kilanioti
Department of Computer Science, University of Cyprus
ekoila01@cs.ucy.ac.cy

CHIPSET TRAINING SCHOOL 2017

BIG DATA PROCESSING IN THE INTERNET OF EVERYTHING ERA

NOVI SAD, SERBIA, 22 SEPTEMBER

Outline

INTRODUCTION AND MOTIVATION

- **Introductory Definitions**
- **Introduction and Motivation**

APPROACH

- **Phase 1: Literature Review on Content Delivery over OSNs**
- **Phase 2: The Social Prefetcher**
- **Phase 3: Parameterization**
- **Phase 4: Predictive Model for Diffusions over OSNs**

DISCUSSIONS AND CONCLUSIONS

- **Large-scale Datasets**
- **OSN Evolution**
- **Semantic Annotation**
- **Mobile CDNs and the Cloud**

Introductory definitions
Online Social Network (OSN)

- ✓Depicted by a directed graph G = (V,E) (social graph)
- ✓Semantics of edges
- ✓Directionality of the edges
- ✓Neighbourship

OSN-aware systems /algorithms / mechanisms: take information extracted from OSNs into consideration:

- the general structural properties of the OSN
- information exchanged over the OSN

Introductory definitions … Information Diffusion and Social Cascades

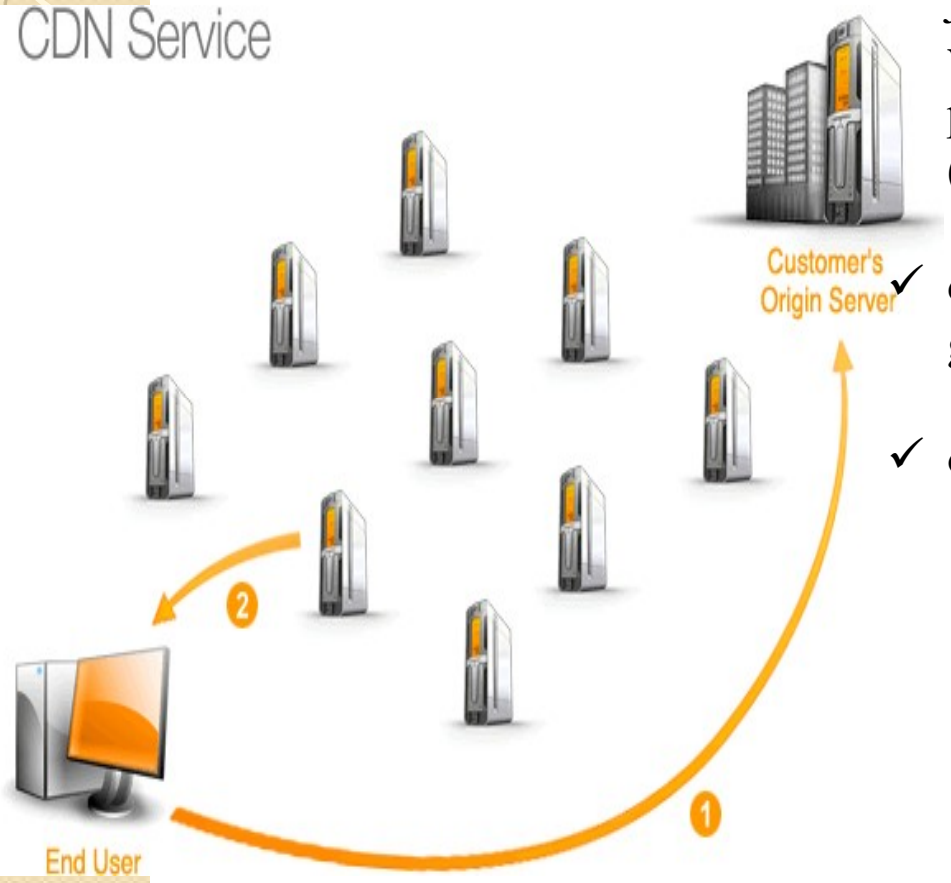
Information diffusion:

a piece of information will become eventually popular or its spread will stop quickly

Social Cascade:

a piece of information is extensively retransmitted over an OSN after its initial publication from an originator user

Introductory definitions…
Content Delivery Networks (CDNs)



- ✓ dynamic replication of data ((images, CSS, javascript files (webpage assets))) in various places of the world as near as possible to the user that consumes it (surrogate servers (2) closer to location)
- ✓ dissimilar in terms of provided services / geographic coverage
- ✓ optimization of their overall efficiency:
  - ➢ automatic detection of the medium (pc/ smartphone / tablet),
  - ➢ optimized management of the browser cache,
  - ➢ server load-balancing,
  - ➢ consideration of specific nature of the content of the media provider (video on demand, live videos, geo-blocked content, etc.)

Introductory definitions…
CDN Copy policies

- **Push-based** proactive prefetching of content to all surrogate servers (minimum response time, maximum copying content cost)
  - typically used for information that is in high demand by users (large files or static assets that don't frequently change as often)
- **Pull-based** content is forwarded to the surrogate server at the moment the user asks for it (minimum copying cost, maximum response time)
  - typically used for personalized information (ideal for small objects with inherent virality and limited duration),
  - Most modern CDNs: MaxCDN, EdgeCast, Amazon CloudFront, BitGravity, Akamai, CDNetworks, CacheFly, ChinaCache, MaxCDN, CDN77, etc. still deploy both pull and push zones, with pulling being the most dominant case.

- **Cooperative:** (to reduce replication and update cost) surrogate servers are cooperating with each other in case of cache misses
  - closest / random / load balancing
  - mapping between content and surrogate servers

- **Uncooperative:** local surrogate server or origin server

Introductory definitions…
CDNs suitable for …

- Sites streaming large video files
- Sites which consist of mainly large media files like image sites
- Sites with known heavy traffic in different countries
- Sites with many tablet and mobile users
- Not for sites that have their main traffic in one geographic area or region

online multimedia streaming providers (e.g. YouTube) rely on CDNs [88]

Introduction and Motivation (1/2)

•**multimedia content delivery technologies**: essential for a wide range of innovative services-multimedia social networks, P2P video streaming, IPTV, interactive online games, cloud multimedia content delivery, content-centric networks

• **Network infrastructures**: **Content Delivery Networks (CDNs)**

delivery of voluminous content:

•proliferation of smartphones

•cheap broadband connections

•free short clip and streaming platforms *(100 hours of video content uploaded in YouTube per minute [14])*

•multiplication over popular **Online Social Networks (OSNs)** *(500 million tweets per day, of which more than 400 tweets per minute include a YouTube link ([10], [7], [39]))*

•video stalling events (i.e. playback interruptions) have a

CDNs need to cope with the **cost-efficient prefetching of bandwidth-intensive content**.

Introduction and Motivation (2/2)

Major CDN issues [92]:

- (i) the *most efficient placement of surrogate servers* (high performance and less infrastructure cost)
- (ii) the *best content diffusion placement* (which content will be copied and to what extent)
- (iii) the *temporal diffusion* (most efficient timing of the content placement)

OSN issues:

- efficient handling of graphs with billions of nodes and edges [161]
  global replication demanded by traditional CDNs becomes expensive
- Long-Tail effect of user-generated content

  not popular enough to be replicated globally, but together the long-tail may get sufficient accesses

Combined issues:

unchanged throughput of the proposed systems / algorithms / policies with the increase in the data input size (social graphs, cascades)

Content staging -Limitations / Caveats

OSNs utilization:

- *lacks experimental evaluation with non-synthetic workloads*
- *ignores storage issues of the infrastructure employed*
- *overlooks matters such as refined topology of the employed data centers*

challenging endeavor: The engineering of general OSN-aware content placement policies over a CDN infrastructure

Aim of the Research Work: "the exploitation of usage patterns found in OSNs for the improvement of user experience through the facilitation of proactive content caching decisions in existent Content Delivery infrastructures"

Outline

INTRODUCTION AND MOTIVATION → **APPROACH** | **DISCUSSIONS AND CONCLUSIONS**

- **Introductory Definitions**
- **Introduction and Motivation**

- **Phase 1: Literature Review on Content Delivery over OSNs**
- **Phase 2: The Social Prefetcher**
- **Phase 3: Parameterization**
- **Phase 4: Predictive Model for Diffusions over OSNs**

- **Large-scale Datasets**
- **OSN Evolution**
- **Semantic Annotation**
- **Mobile CDNs and the Cloud**

Research work phases

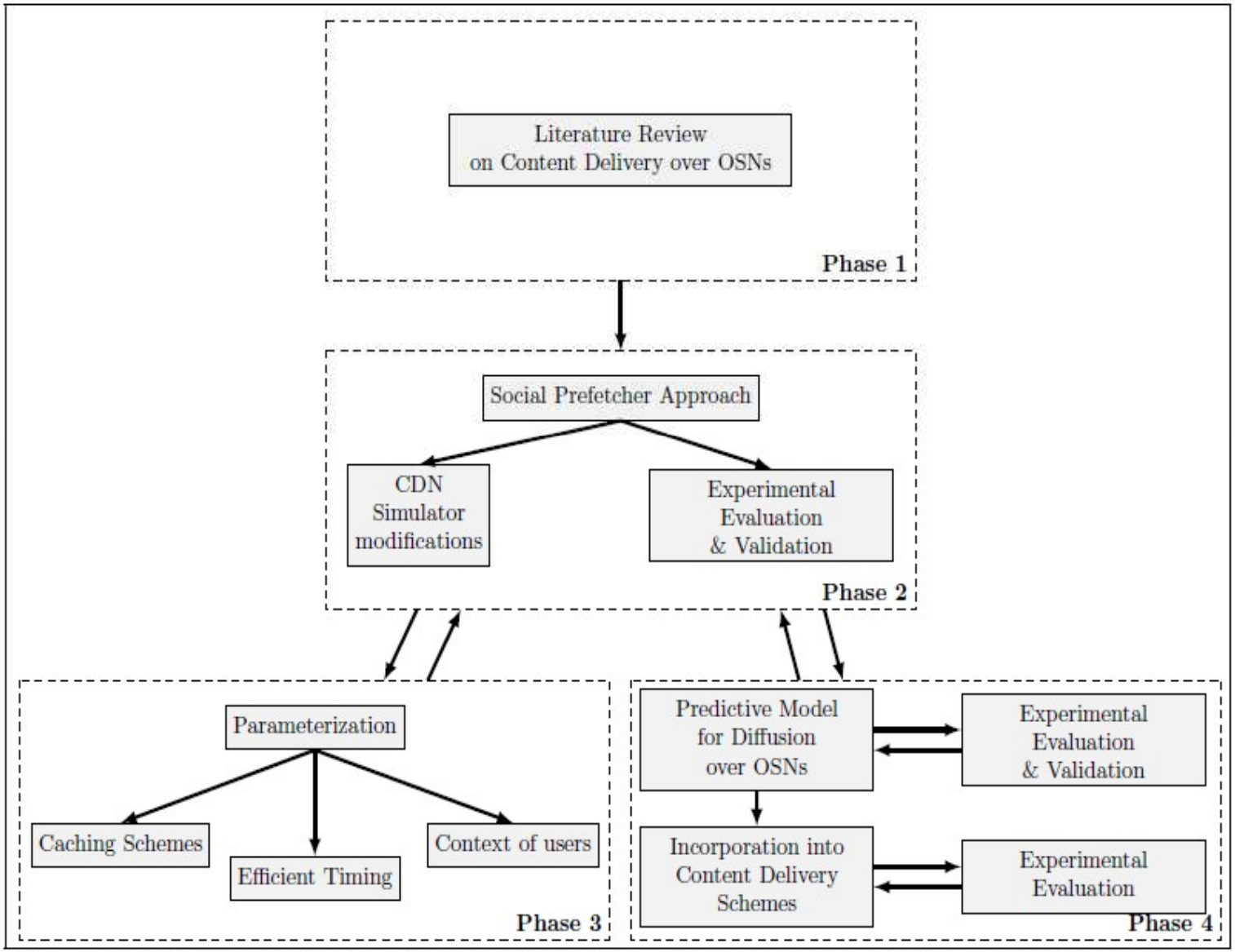


Fig. 1: Phases of Research Work

Related Work – Taxonomy of Content Delivery over OSNs [94]

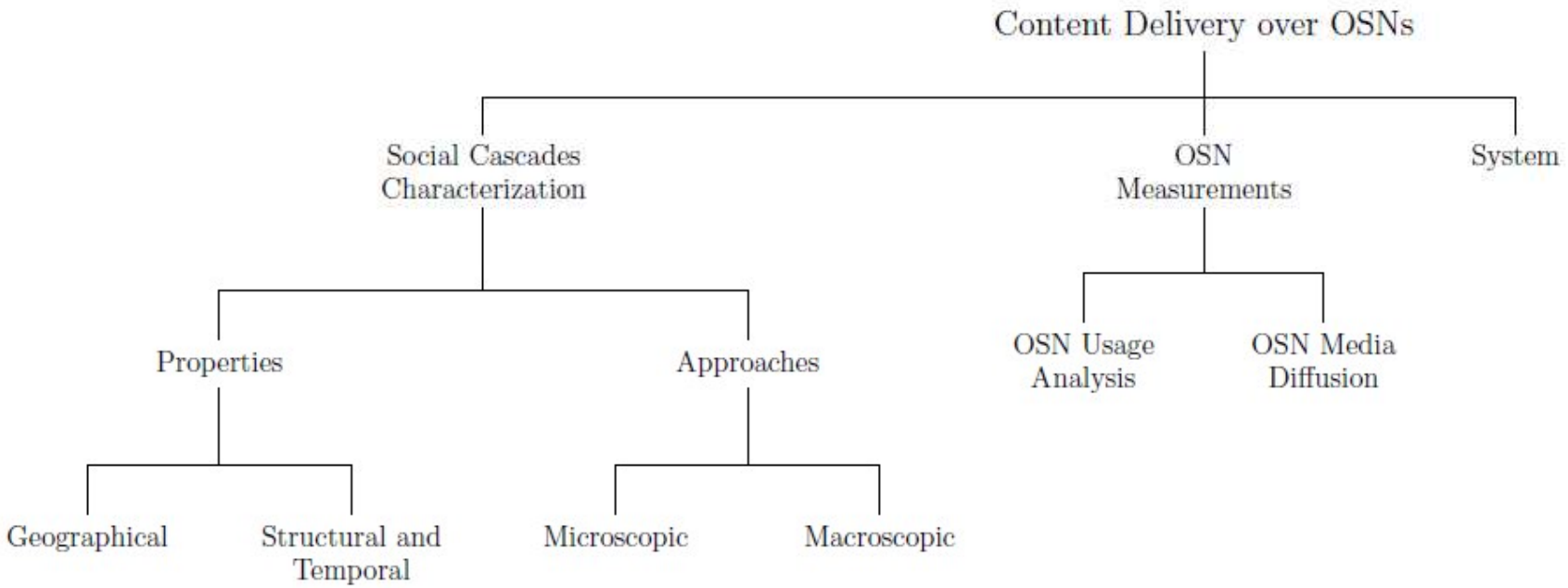
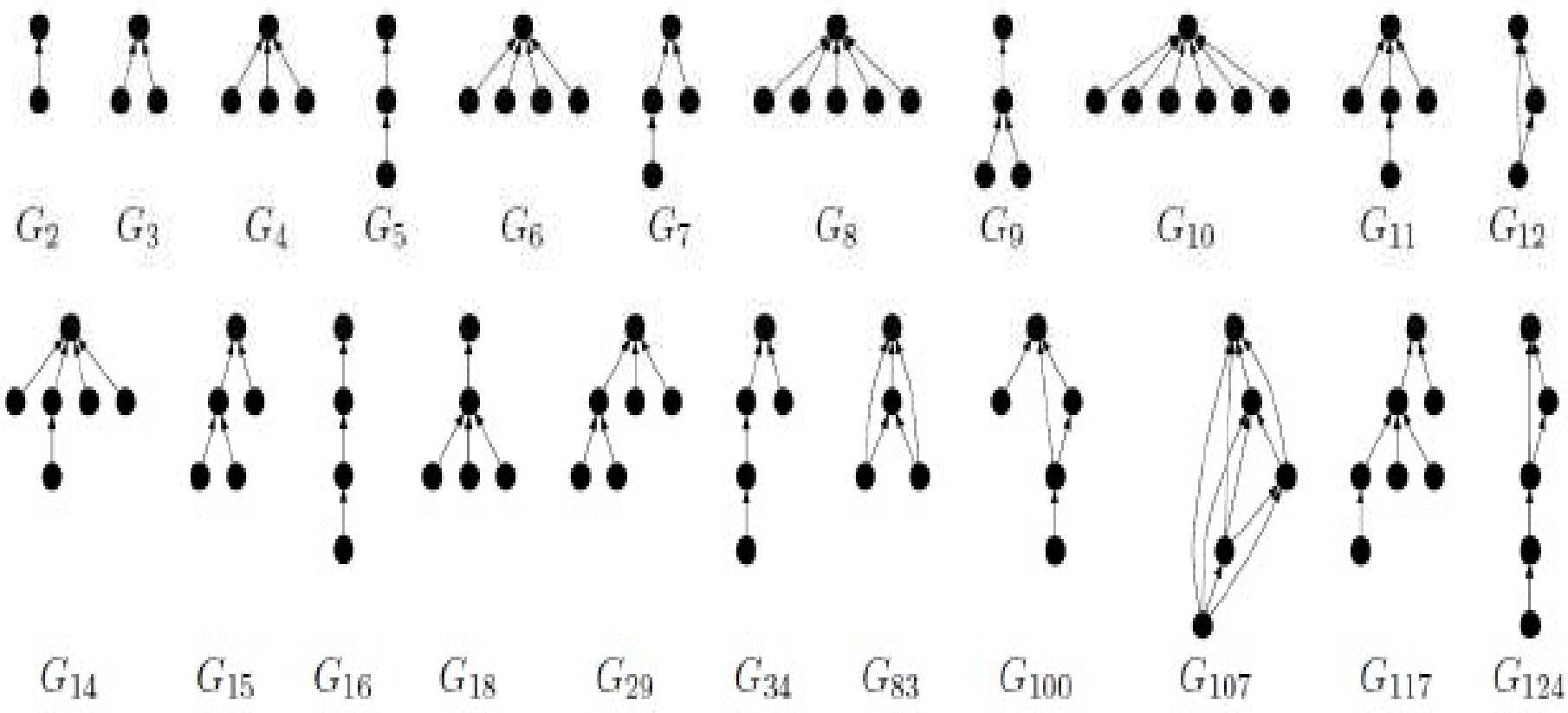


Fig. 2: A taxonomy of Content Delivery over OSNs.

[94] I. Kilanioti, C. Georgiou, and G. Pallis, "On the impact of online social networks in content delivery," in Advanced Content Delivery and Streaming in the Cloud, M. Pathan, R. Sitaraman, and D. Robinson, Eds. Wiley, 2014.

Metrics for characterization of cascades - Common cascades



Ordering of common shapes of cascades in the blogosphere [73] by frequency, with r the frequency ranking of Gr.

Metrics for characterization of social cascades

- Geographic (geodiversity, georange [138])
- Structural (size, length [27])
- Temporal (time delay between consecutive steps [138], time duration, rate of the cascade [44])

Approaches:

- Microscopic (Watts [157])
- Macroscopic (Kleinberg and Easley [60], Ver Steeg et al. [145])
- Hybrid (Dave et al. [52])

Bandwidth-intensive media content and Social Networks – Measurement Studies on OSNs

[19],[85],[98],[101],[105],[117],[159]:

- power-law $E(t) \propto V(t)^a \quad a \in (1,2)$ scale-free $P(k) \sim k^{-\gamma}$
- in-degree matches out-degree,
- average distances are lower, clustering coefficients higher than those of the web graph (clustered 10.000 more times than random graphs, 5–50 times more than random power-law graphs) $C_i = \frac{2|(v,w)|, (i,v), (i,w), (v,w) \in E}{k_i(k_i - 1)}$

- giant component (dense core of shrinking diameter [106]) / middle region(various isolated communities interacting with one another but not with the overall network)/ singletons [98]

- temporal evolution (densification power-law, shrinking diameters)

- OSN user workloads: clickstream model [34]
  - browsing most dominant behavior (92%),
  - social cascade effect (bandwidth-intensive-media found through a 1-hop friend, 80%)

Impact of bandwidth-intensive media content diffusion over OSNs

- [118], [43], [66], [71], [127]:YouTube traffic with emphasis on the file size, bitrate, usage patterns and popularity
- [71]: similarities between traditional Web and media streaming workloads
- [47]:YouTube videos, small-world characteristics
- [65], [164], [42]: Long-tail effect for YouTube and VOD systems
- [163]:temporal variation of popularity of content in OSNs

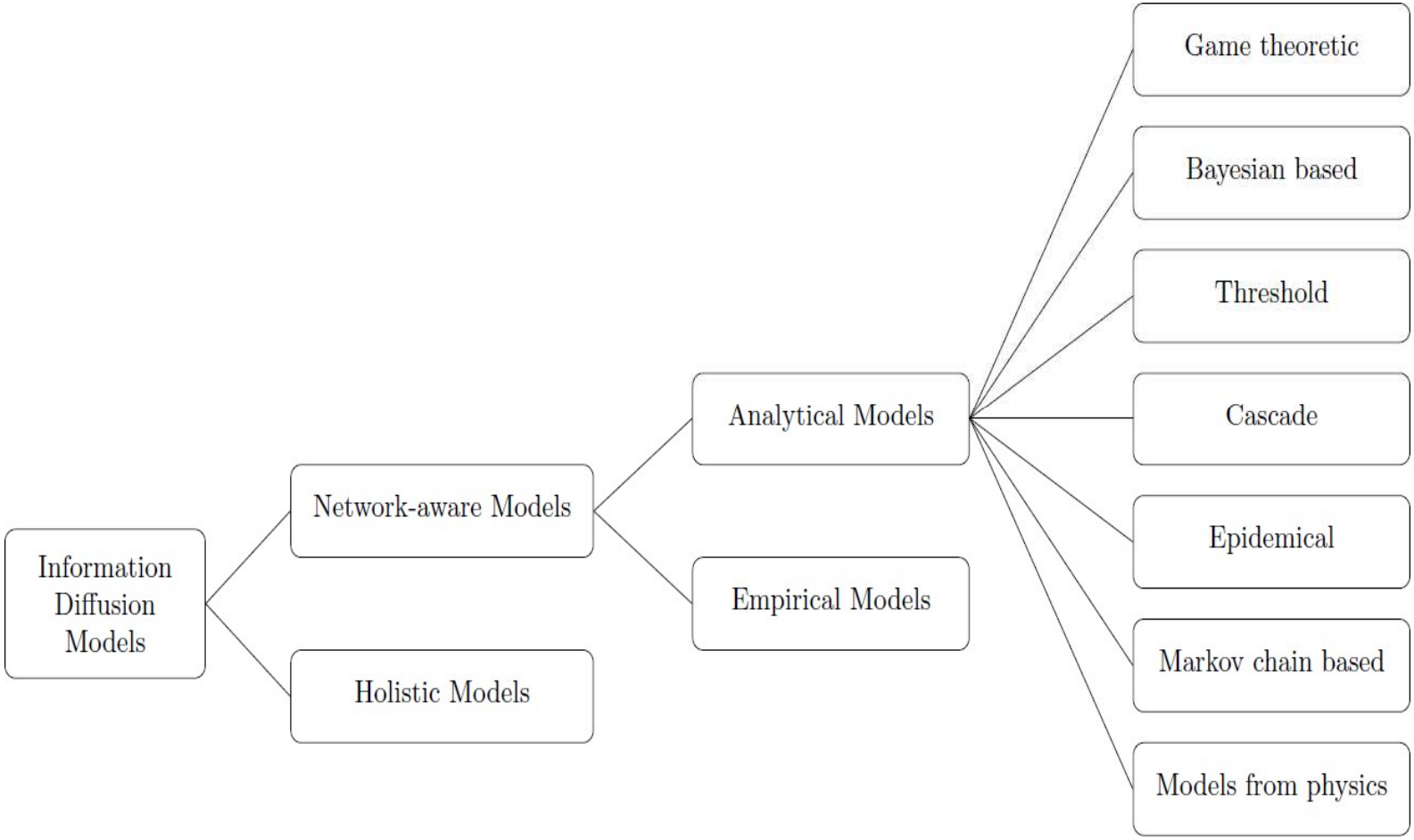
Applications and techniques (1/2)

- Buzztraq [137]: generation of hints for replica placement based on the users' friends' location and number
  - outperforms location based placement (geographical location of recent users)
  - server bandwidth and storage constraints: ignored
  - social cascade is indirectly analyzed via a third-party page (access to media and social profile)

Applications and techniques (2/2)

- [166]: web based scheme for caching using the access patterns of friends within the same Internet Service Provider (ISP) with a drop-in component
  - users protected with k-anonymity
- [81]: logical addressing scheme technique for putting together in the disk blocks containing data from friends
  - greedy heuristic that finds a layout for the users within the communities
  - organizes the different communities on the disk by considering inter-community tie strength
- [82]: content locality (induced by the related videos feature) and geographic locality are in fact correlated
- [138]: proof-of-concept geographic model of CDN
  - "social cascades tend not to expand geographically"

Models of information diffusion



SNA tools

Name	Purpose	Built on	GUI	Mode
SNAP	analytical	C++	available through graphical front-end NodeXL	single work-station
NetMiner	analytical	Python	yes	single work-station
igraph	analytical	R, Python	visualization capabilities	single work-station
NetworkX	analytical	Python	visualization capabilities	single work-station
NetEvViz	visualization of temporal differences	C#	yes	single work-station
XRime	analytical	Java	no	MapReduce

Tools for CDN simulation

Name	Purpose	Built on	GUI
CDNSim	solely for CDNs	OMNET++, INET	yes
NS-2	general purpose	C++, simulation scenarios in Object Tcl	no
NS-3	general purpose	C++, python	no
PlanetLab	general purpose	Linux vserver as node provisioning mechanism, migrating to LXC - the implementation of container-based virtualization in the Linux kernel	yes

Graph tools

Name	Purpose	Built-on	Features
PEGASUS	graph storage, graph mining	Java	indexing, inference, spectral analysis, node-centralized computation exclusively
GBASE	graph storage, graph mining	MapRedude framework	node-centralized or edge-centralized computation
Mondal Deshpande approach	dynamic replication of nodes based on read-write frequencies	Apache Couch-DB key-value store	exploits clustering
Facebook Corona	improvement of Apache Hadoop scalability	Java	separate central cluster manager and multiple job trackers, task scheduling in push model

Phase 2 – The Social Prefetcher:

<u>Design</u>, experimentally <u>evaluate</u> and <u>validate</u>

- an efficient copying <u>algorithm</u>
- the accompanying <u>framework</u>

a dynamic mechanism of preactive copying based on demand prediction in social networks to a CDN infrastructure

<u>**Major CDN issues** [92]:</u>

- ❑(i) the *most efficient placement of surrogate servers*
- ❑(ii) the *best content diffusion placement* *(which content will be copied, local/global replication extent)*
- ❑(iii) the *temporal diffusion*.

Challenges…

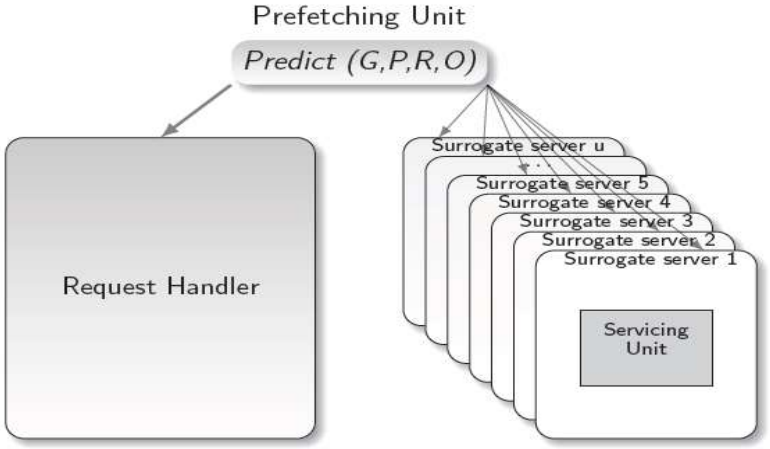
- ▪efficient handling of graphs with billions of nodes and edges
- ▪efficient handling of long-tail UGC (virality, localization)
- ▪real datasets for study of cascades
- ▪data placement, replication and distribution for a large variety of resource types and media formats
- ▪blackbox treatment of CDN policies/ need for participation of third users

[92] I. Kilanioti, “Improving multimedia content delivery via augmentation with social information. The Social Prefetcher approach.” Multimedia, IEEE Transactions on, vol. 17, no. 9, pp. 1460–1470, 2015. [Online]. Available: http://goo.gl/x81Xv1

Phase 2: The Social Prefetcher [92]

Table 1: Notation Overview

$G(V, E)$	Graph representing the social network
$V = \{V_0, V_1, ..., V_n\}$	Nodes representing the social network users
$E = \{E_{00}, E_{01}, ..., E_{0n}, ..., E_{nn}\}$	Edges representing the social network connections, where E_{ij} stands for friendship between i and j
$R = \{r_1, r_2, ..., r_\tau\}$ $r_i \subset V$	Regions set
$N = \{n_1, n_2, ..., n_u\}$	The surrogate servers set. Every surrogate server belongs to a region r_i
$C_i, i \in N$	Capacity of surrogate server i in bytes
$O = \{o_1, o_2, ..., o_w\}$	Objects set (videos), denoting the objects users can ask for and share
$S_i, i \in O$	size of object i in bytes
Π_i	popularity of object i, $i \in O$
$q_i = \{t, V_\psi, o_x\}$	Request i, consists of a timestamp, the id of the user that asked for the object, and the object id
$P = \{p_{01}, p_{02}, ..., p_{nw}\}$	User posts in the social network, where p_{ij} denotes that node i has shared object j in the social network
$Q = \{q_1, q_2, ..., q_\zeta\}$	Object requests from page containing the media objects, where q_i denotes a request for an object of set O



$$Put(n_i, Predict(G, P, R, O))$$

$$\frac{Q_{hit}}{Q_{total}}$$

$$\sum_{\forall i \in O} S_i f_{ik} \leq C_k$$

$$f_{ik} = \begin{cases} 1 & \text{if object } i \text{ exists in the cache of surrogate server } k \\ 0 & \text{if object does not exist} \end{cases}$$

Applied heuristics

- Users more influenced [92] :
  - by geographically close friends → "geographic zones"
  - moreover by mutual followers

- Social cascades: short duration [138], [174]
  - percentage of cascades proceeding for days not directly attributed to the influence that a social contact exerts (video in the user newsfeed)
  - threshold for the cascade effect :
    - *24 hours/ 48 hours/ threshold covering all requests/ indicatively <24 hours*

New requests in the CDN

Algorithm for every new request <timestamp, V, o> in the Content Delivery Network:

```
1  if o.timestamp == 0 then
2  |   o.timer = 0
3  |   o.timestamp = request_timestamp
4  else if o.timestamp != 0 then
5  |   o.timer = o.timer + (request_timestamp - o.timestamp)
6  |   o.timestamp = request_timestamp
7  end
8  if o.timer > time_threshold then
9  |   o.timer = 0
10 |   o.timestamp = 0
11 else if o.timer < time_threshold AND user.authority_score >
   authority_threshold then
12 |   copy object o to surrogate that serves user's V timezone
13 |   for each user y that follows user V
14 |   find surrogate server S that serves y's timezone
15 |   copy object o to S
16 else if o.timer < time_threshold then
17 |   copy object o to surrogate that serves user's V timezone
18 |   copy object o to surrogates S that sub-policy I OR sub-policy II OR
   |   sub-policy III decides
19 end
```

Subpolicies for local copying

Subpolicy I for local copy

```
1 find X timezones where (user V has mutual followers AND they are closer to
  user's V timezone)
2 foreach timezone that belongs to X do
3 |   find surrogate server S that serves timezone
4 |   Copy object o to S
5 end
```

Subpolicy II for local copy

```
1 find X timezones where (user V has mutual followers AND they are closer to
  user's V timezone)
2 find the L ⊆ X that (belong to X AND have the highest lobby-index score)
3 foreach timezone that belongs to L do
4 |   find surrogate server S that serves timezone
5 |   Copy object o to S
6 end
```

Subpolicy III for local copy

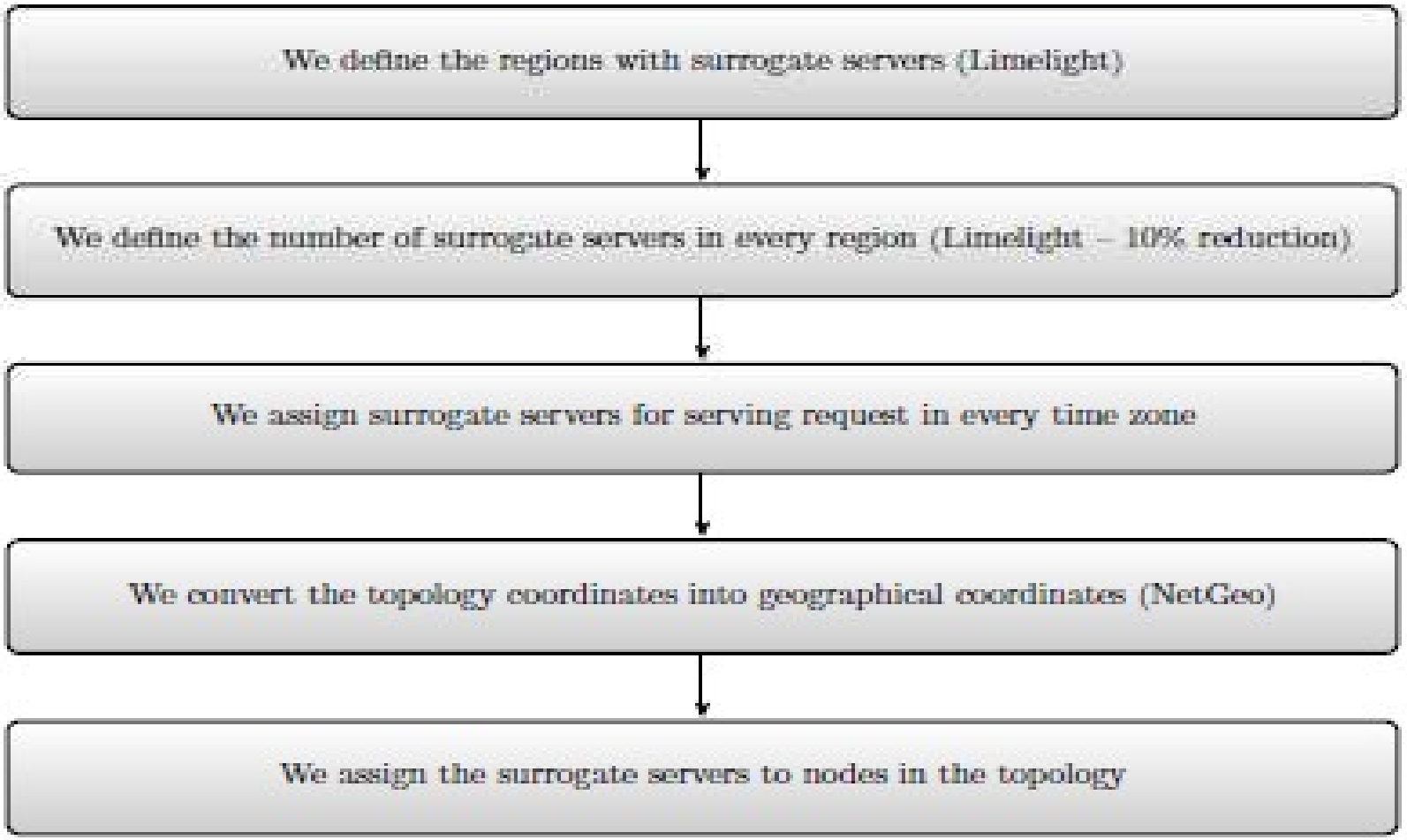
```
1 find X timezones where (user V has mutual followers AND they are closer to
  user's V timezone)
2 find the H ⊆ X that (belong to X AND have the highest HITS score)
3 foreach timezone that belongs to Y do
4 |   find surrogate server S that serves timezone
5 |   Copy object o to S
6 end
```

For every new object in the server

Algorithm for every new object o in the surrogate server S

```
1  if o.size + current_cache_size ≤ total_cache_size then
2  |  copy object o in surrogate S' cache
3  else if o.size + current_cache_size > total_cache_size then
4  |  while o.size + current_cache_size > total_cache_size do
5  |  |  foreach object q in cache do
6  |  |  |  if (current_timestamp - q.timestamp) + q.timer > time_threshold
   |  |  |  then
7  |  |  |  |  copy q in CandidateList
8  |  |  |  end
9  |  |  |  if CandidateList.size>0  AND  CandidateList.size !=
   |  |  |  total_cache_size then
10 |  |  |  |  find q that q.timestamp is maximum and delete it
11 |  |  |  else if CandidateList.size==0 OR
   |  |  |  CandidateList.size==total_cache_size then
12 |  |  |  |  Use LRU to delete any object o ∈ O
13 |  |  |  end
14 |  |  end
15 |  end
16 |  Put object o to surrogate's S cache
17 end
```

CDN proof-of-concept setup Methodology



Evaluation

City	Servers	City	Servers
Washington DC	55	Toronto	12
New York	43	Amsterdam	20
Atlanta	11	London	30
Miami	11	Frankfurt	31
Chicago	37	Paris	12
Dallas	19	Moscow	10
Los Angeles	52	Hong Kong	8
San Jose	37	Tokyo	12
Seattle	15	Changi	5
Phoenix	3	Sydney	1

✓162 zones, as the 142 zones of Twitter include generic characterizations, e.g., Eastern Time and Central Time

3 Pacific Time, 14 none, 1 international, 1 westafrica, 1 Midatlantic, 4 Eastern, 3 Central Time

Dataset and Experimentation Setup

- ✓ Twitter dataset containing geographic locations, follower lists and tweets for 37 M users
- ✓ spreading of more than 1M YouTube videos over this network
- ✓ a corpus of more than 2 B messages and
- ✓ approximately 1.3 M single messages with an extracted video URL

User	
Id	User id
Verified	If user has verified email
followers_count	Number of user's followers
Protected	If user's information is private
listed_count	How many tweets refer the user
statuses_count	How many tweets the user has published
friends_count	How many users the user follows
Location	Explicit location of the user
geo_enabled	If the service denoting the user location along with tweet is enabled
Lang	User language
favourites_count	How many tweets user has added to favourites
created_at	Time of creation
time_zone	Timezone of the user

Tweet	
Id	Tweet id
Text	Tweet content
created_at	Time of creation
Retweeted	If it is retweet
in_reply_to_status_id	status id of the tweet to which it replies
in_reply_to_user_id	User id of the tweet to which it replies
urls	Urls included
retweet_count	Number of retweets

- ✓ 330 experiments (55 per (time threshold & centrality metric): *all possible combinations for X=10 closest geographic zones and Y<X zones with highest centrality*)
  - ✓ m1.xlarge AWS EC2 instance (ca. 6 hours) per experiment,
  - ✓ UCY VPS
- ✓ AWS Elastic MapReduce for Graph analysis (centrality computation)

Metrics

- Mean Response Time: how fast a client is satisfied

$\frac{\sum_{i=0}^{N-1} t_i}{N}$ N number of satisfied requests, t_i response time of i-th request

- Hit Ratio: percentage of the client-to-CDN requests that resulted in a cache hit (high values: high-quality content placement of the surrogate servers)
- Active servers: servers being active serving clients
- Mean Surrogate Servers Utility: number of bytes of the served content against the number of bytes of the pulled content (from the origin server or other surrogate servers)

Experimental results (1/4)

- Impact of time threshold duration

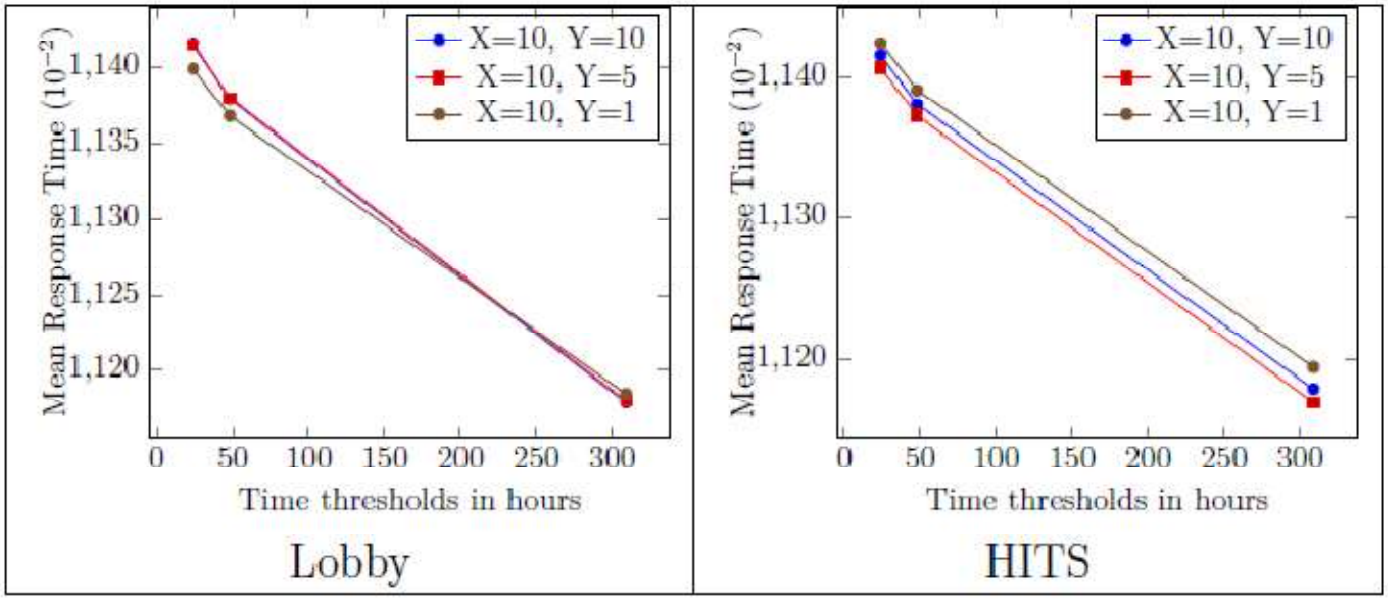


Table 3: Effect of time threshold duration on mean response time, for X closest timezones with mutual followers and Y timezones with the highest metric, where copying is ultimately done

Experimental results (2/4)

- Impact of number of zones (MRT)

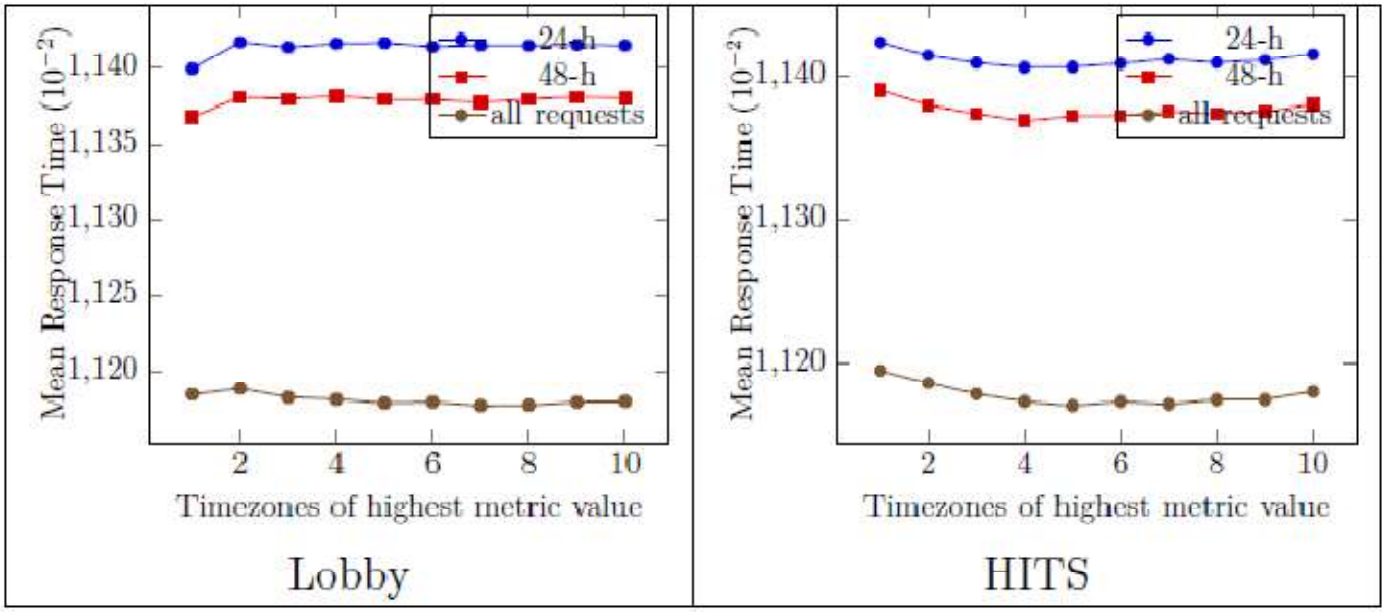


Table 7: Effect of time zones used as Y on mean response time, for $X = 10$ closest timezones with mutual followers

Experimental results (3/4)

- Impact of influence measurement metric (MRT)

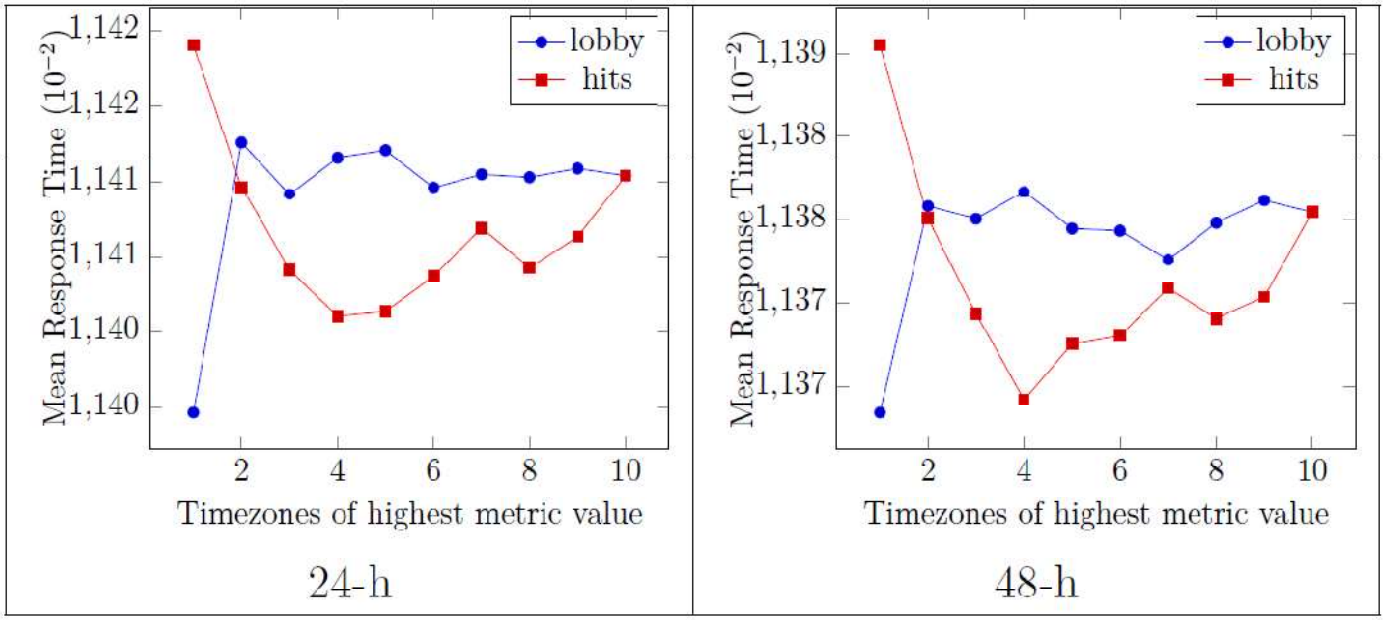
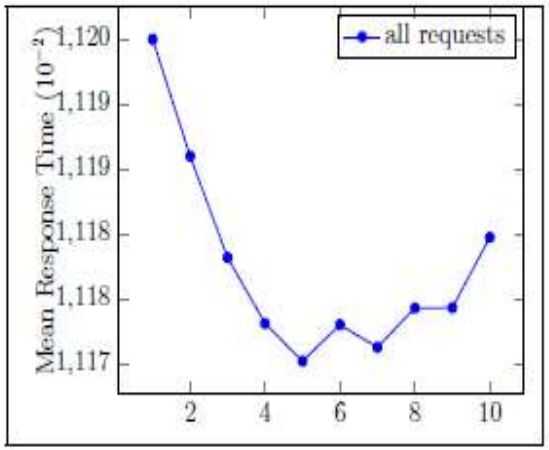


Table 9: Effect of influence measurement on mean response time, for $X = 10$ closest timezones with mutual followers

Experimental results (4/4)



Social prefetcher Improvement up to 40%	**Buzztraq** Improvement 10 to 40% compared to plain LOCATION BASED PLACEMENT
Static 1.846021ms	Plain LBP (k=3) 1.732023s
MRT 1.117026ms	MRT 1.395011ms

Table 8: Mean response time, for X=10 closest timezones with mutual followers and all possible Y values, $Y \in [1, 10]$

- Significant improvement over respective improvement (30%) in pull-based methods employed by most CDNs
- Up to 40% improvement over static policy each time
- Refined data centers topology, storage issues employed methods do not consider

Phase 3 – Parameterization:

Parameterize with

- **caching schemes** variations for the distributed infrastructures the CDNs deploy
- **temporal factors** related to the most efficient timing of the content placement
- other **contextual information** of the OSN and the media platform

[93] I. Kilanioti and G. A. Papadopoulos, "Socially-aware multimedia content delivery for the cloud," in 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC), Dec 2015, pp. 300–309. [Online]. Available: http://goo.gl/gsjh0E

[168] I. Kilanioti and G. A. Papadopoulos, "Delivering social multimedia content with scalability," in Resource Management for Big Data Platforms: Algorithms, Modelling and High-Performance Computing Techniques, Springer Computer Communications and Networks Series, Eds. F. Pop, J. Kolodjiez, B. D. Martino, Springer, 2016.

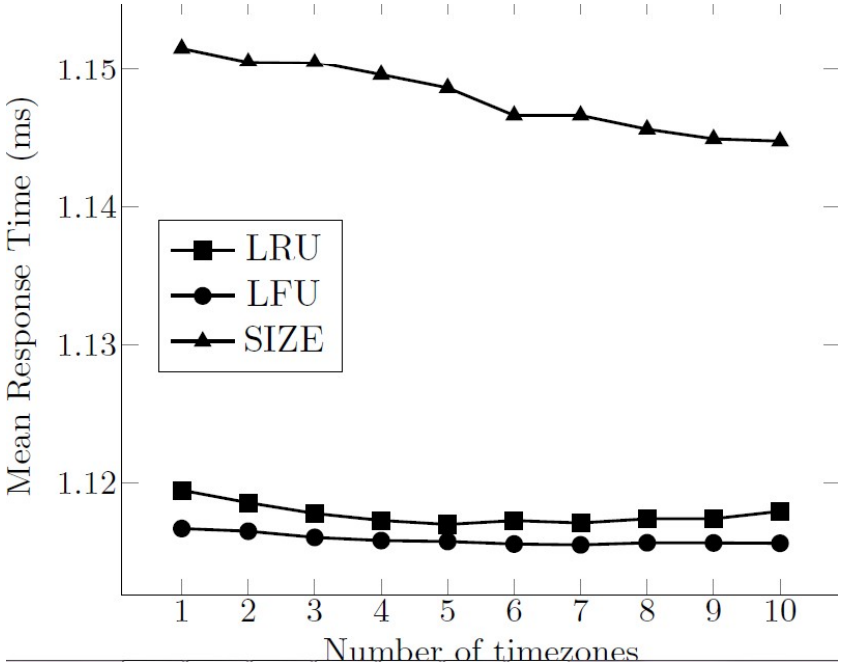
CACHING SCHEMES

- LRU
- LFU
- Size-adjusted LRU/ SIZE

Name	Primary Key	Secondary Key
LRU	Time Since Last Access	
LFU	Frequency of Access	
SIZE	Size	Time Since Last Access

$$S_1 \cdot \Delta T_{1k} \leq S_2 \cdot \Delta T_{2k} \leq \ldots \leq S_{|C(k)|} \cdot \Delta T_{|C(k)|k}$$

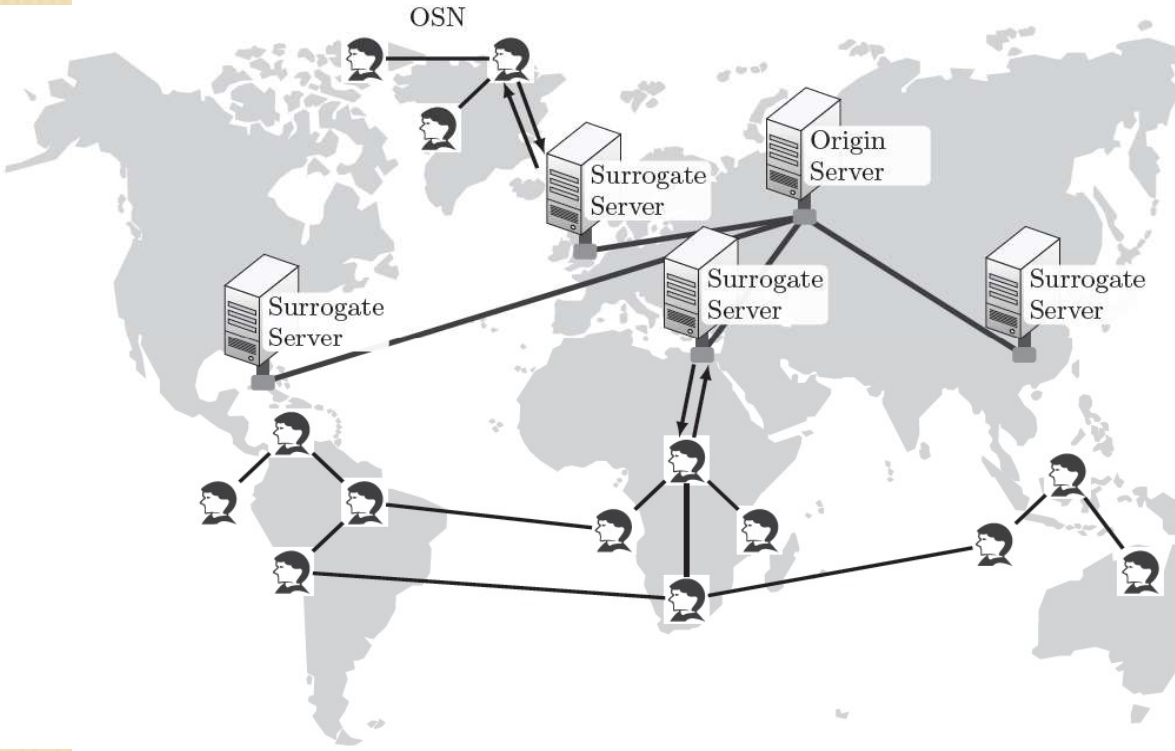
Si: size of object i
C(k): set of objects in cache at k-th iteration
ΔT_{ik} : time since last access of object i (k-th iteration)



Selection of caching schemes among
-mLRU,
-scoring based SC caching algorithm,
-Cache Management based on Temporal Pattern Solicitation (CMTPS) algorithm etc.
based on criteria of:
-Time complexity
-Ease of implementation

Phase 3 – Parameterization [93]
Why is our Approach Necessary: An Example

- Bob(UK): assigned to the local CDN servers of an OSN service



- Bob logs into the OSN and posts a video that he wants to share
- Aggregated over all users, pushing can lead to traffic congestion (content may not be consumed)
- intensified problem of caching with unique friends per area (Alice in Athens)

Copying under conditions…

- Contextual conditions (variation 1)
  - content with high viewership within the media service
- Temporal conditions: at the time window that signifies (variation 2)
  - a non-peak-time for the upload in UK area and
  - a non-peak-time for the download in Athens area

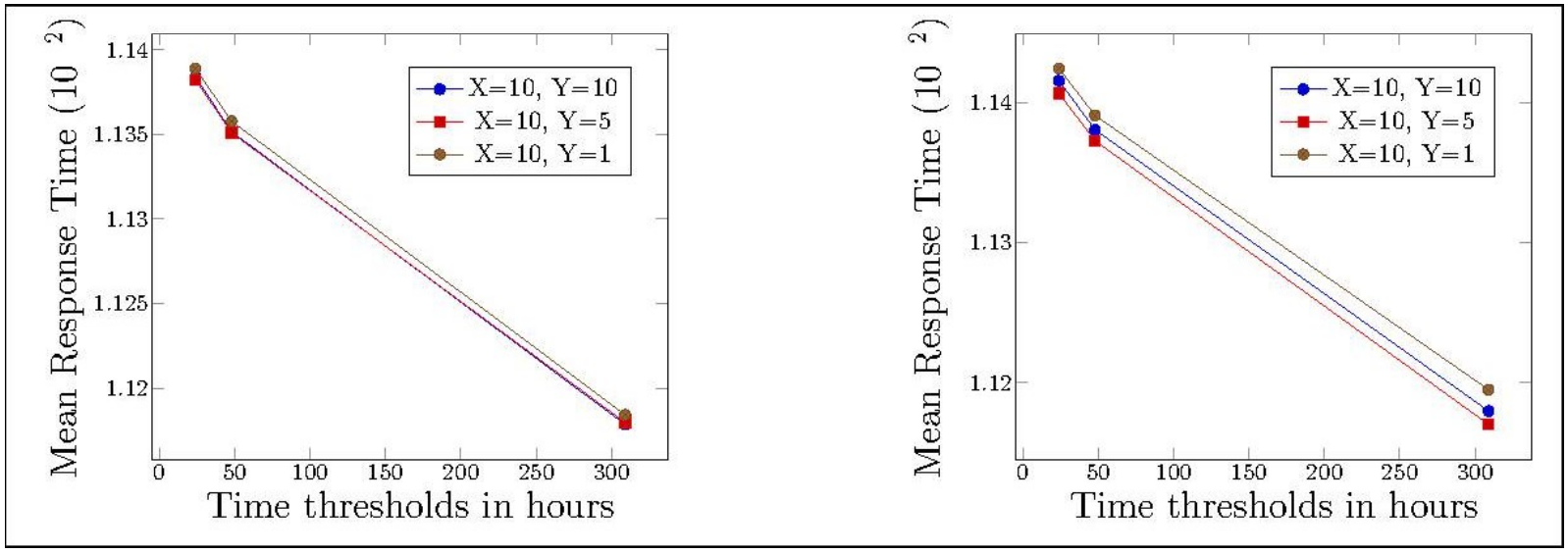
copied to geographically close zones where the user has mutual friends with high influence impact

- HENCE:
- -smaller response times for the content to be consumed (users)
- -lower bandwidth costs (OSN provider)

Impact of Time Threshold Duration

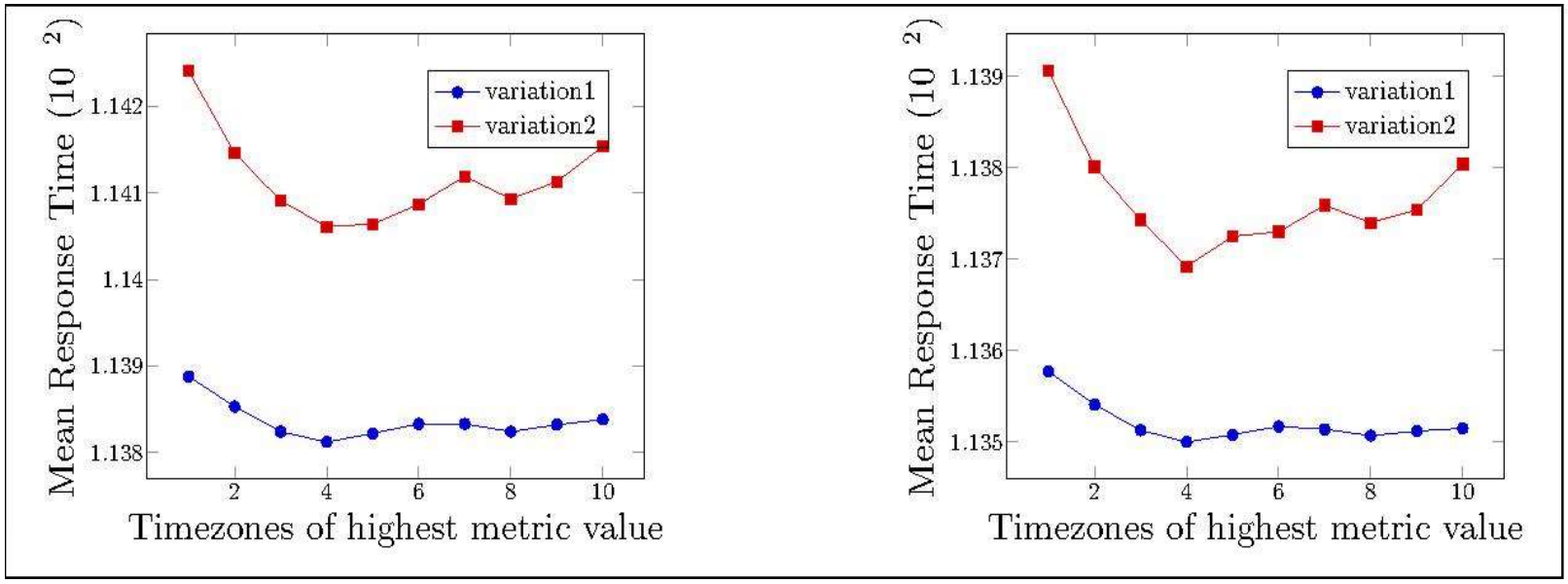
on Mean Response Time: As the time threshold increases from 24 to 48 h and to hours covering the entire set of requests, we observe that the mean response time decreases steadily.

indicative values for the 10 closest zones of mutual followers and varying subsets of 1, 5 and 10 zones with the highest influence metric, respectively, for both variations



Variation 1 Variation 2

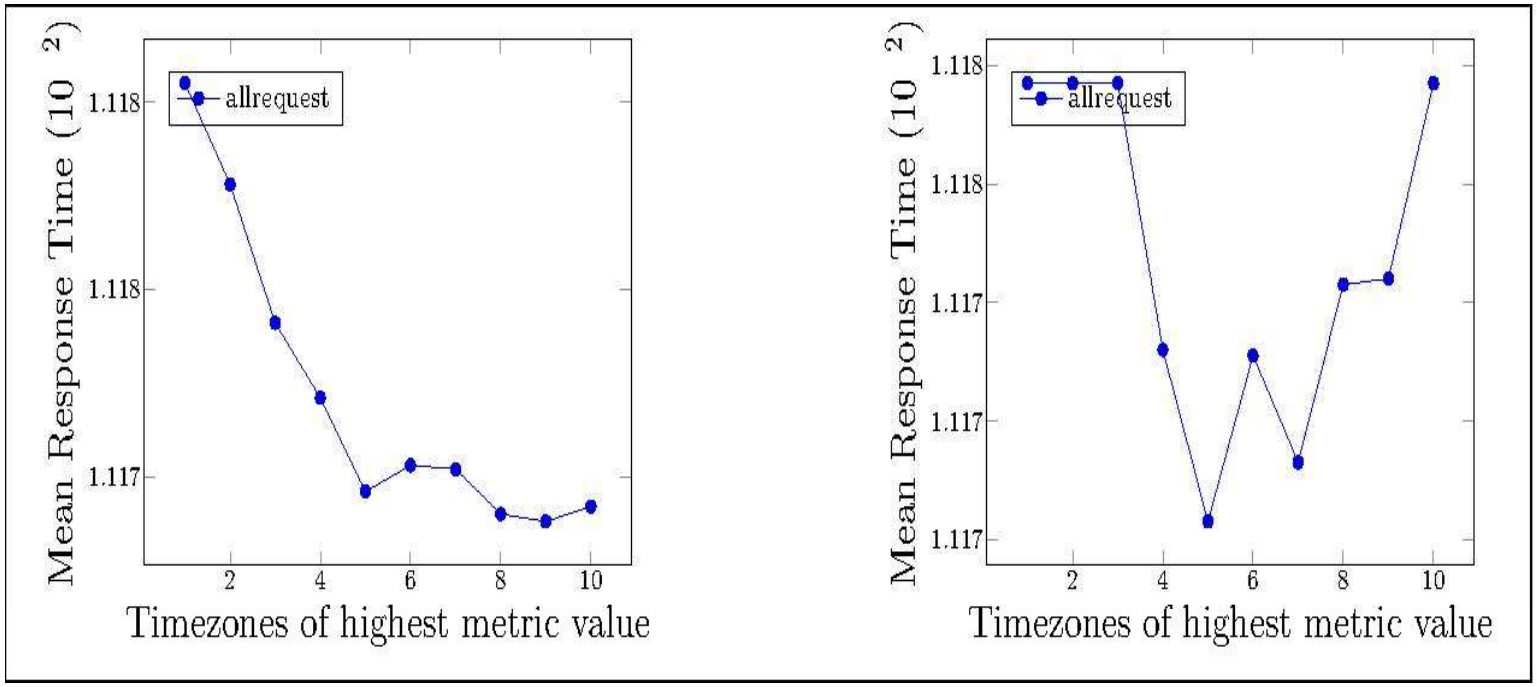
Impact of zones number



24-h 48-h

<u>on Mean Response Time:</u>

- ✓ trade-off between the reduction of the response time and the cost of copying
- ✓ switch point with approximately 4 zones out of the 10 used (for a fixed number of closest zones with mutual followers)
- ✓ slight increase in the mean response time attributed to the delay for copying content to surrogate servers



Variation 1 Variation 2

Mean response time for X=10 closest zones with mutual followers and all possible Y values, Y ∈ [1,10] for (i)Variation-1 and (ii)Variation-2

- ✓ cost per copy: related to the number of hops among the client and the server where copying is likely to be made (Put function)

	Mean response time (Avg. 10^-2 sec.)	Hit ratio (Avg. %)	Active servers	Mean utility (Avg. %)
Variation-1 - 24-h	1.1383	32.81	326	96.01
Variation-1 - 48-h	1.1352	33.08	326	96.01
Variation-1 - all-h	1.1172	34.58	325	96.04
Variation-2 - 24-h	1.1411	32.13	325	95.98
Variation-2 - 48-h	1.1376	32.43	326	96.00
Variation-2 - all-h	1.1174	34.38	324	96.03
Social Prefetcher 24-h	1.1412	32.12	325	95.98
Social Prefetcher 48-h	1.1377	32.42	326	96.00
Social Prefetcher all-h	1.1181	34.16	325	96.01

- ✓ lowest mean response times when time threshold covers all requests
- ✓ better performance in terms of mean response times and hit ratios achieved for the Variation-1
- ✓ both variations perform better than the Social Prefetcher approach (bare implementation without the variations) [92]

Phase 4 – Predictive Model for Diffusion over OSNs:

•merge user-centric data from OSN with video-centric data from media platform

•investigate ties between predictability of video sharing and the social context of video uploaders

•develop and validate accurate model to predict future popularity of a video resource given features of the underlying network of its initial sharer

•incorporate it into an OSN-aware mechanism for content delivery & experimentally evaluate improvement of the user experience

[167] I. Kilanioti and G. A. Papadopoulos, "Efficient content delivery through popularity forecasting on social media," in Proceedings of the 7th IEEE International Conference on Information, Intelligence, Systems and Applications, IISA 2016, Chalkidiki, Greece, July 13-15, 2016, pp. 13–19.

[169]] I. Kilanioti and G. A. Papadopoulos, "Predicting video virality on Twitter," in Resource Management for Big Data Platforms: Algorithms, Modelling and High-Performance ComputingTechniques, Springer Computer Communications and Networks Series, Eds. F. Pop, J. Kolodjiez, B. D. Martino, Springer, 2016.

Related Work

Prediction of social virality:

What is predicted…

- amount of aggregate activities (e.g., aggregate daily hashtag use)
- user-level behaviour (retransmission of a specific tweet/URL)
- growth of the cascade size

Duration of prediction study…

- specific time-window
- entire cascade duration

Approach….

- Feature-based methods (content, temporal etc. features)
  - learning algorithms schemes: simple regression analysis, regression trees, content-based methods, binary classification, etc.
- Time-series analysis works

TABLE 1. NOTATION OVERVIEW

$G(t) = (V(t), E(t))$	graph G at time t of V vertices and E edges
A_{u2v}	number of actions where u influenced v
$\widehat{A_{u2v}}$	predicted output
M	total number of predicted values
α, β, γ	coefficients of feature set variables
U	vector of YouTube interests of user u
V	vector of Twitter interests of user v
	Features Set
$Score(u,t)$	Score of node u at time t
$dScore = dScore(u,t)/dt$	derivative of Score of node u at time t
$content\_dist$	content distance

$$A_{u2v} = \alpha \times Score(u,t) + \beta \times \frac{dScore(u,t)}{dt} + \gamma \times content\_dist \tag{1}$$

$$\sqrt{\frac{1}{M}\sum_{i=1}^{M}(\widehat{A_{u2v}} - A_{u2v})^2} \tag{2}$$

Dataset

- Twitter
  - retweet mechanism enables users to propagate information across multiple hops in the network
- Analysis of user interests [1] against directory information from http://wefollow.com
- Variety of features extracted:
  - number of users' tweets,
  - fraction of tweets that are retweets,
  - the fraction of tweets containing URLs, etc.
- Sharing events in the dataset: tweets containing a valid YouTube video ID (category, Freebase topics and timestamp)
- Aggregated features of YouTube videos shared:
  - average view count
  - median inter-event time between video upload and sharing, etc.
- Dataset augmentation with Tweet content information
  - for 15 M. video sharing events
  - followership information of the 87K Twitter users

n: number of followers,
b: average of number of a user's followers / number of users he follows (catering for users with reciprocal followership),
e: average number of retweets X number of user's tweets (effect of a user's tweet)

$$Score = \log\left(n + \left(\left(\frac{b}{100}\right) \times n\right) + e\right) \qquad (3)$$

$$content\_dist = 1 - \frac{U \cdot V}{\|U\|\|V\|} \qquad (4)$$

Score calculation, Content distance

Experimental Evaluation

Selection of predictors among:

- number of distinct users retweeted,
- fraction of the user tweets that were retweeted,
- average number of friends of friends,
- average number of followers of friends,
- number of YouTube videos shared,
- account creation time,
- number of views of a video, etc.,

TABLE 2. REGRESSION RESULTS WITHOUT OUTLIERS (I)

Dep. Variable	A_{u2v}	R-squared	0.629
Model	OLS	Adj. R-squared	0.629
Method	Least Squares	F-statistic	3.072e+04
Prob (F-statistic)	0.00	Log-Likelihood	13947.
No. Observations	54473	AIC	-2.789e+04
Df Residuals	54470	BIC	-2.786e+04
Df Model	3	Covariance Type	nonrobust

TABLE 3. REGRESSION RESULTS WITHOUT OUTLIERS (II)

	coef	std err	t	$P > \|t\|$	95%	Conf.Int.
Score	0.1460	0.001	145.244	0.000	0.144	0.148
dScore	0.0200	0.001	25.819	0.000	0.018	0.022
con_dist	0.1656	0.003	65.690	0.000	0.161	0.171

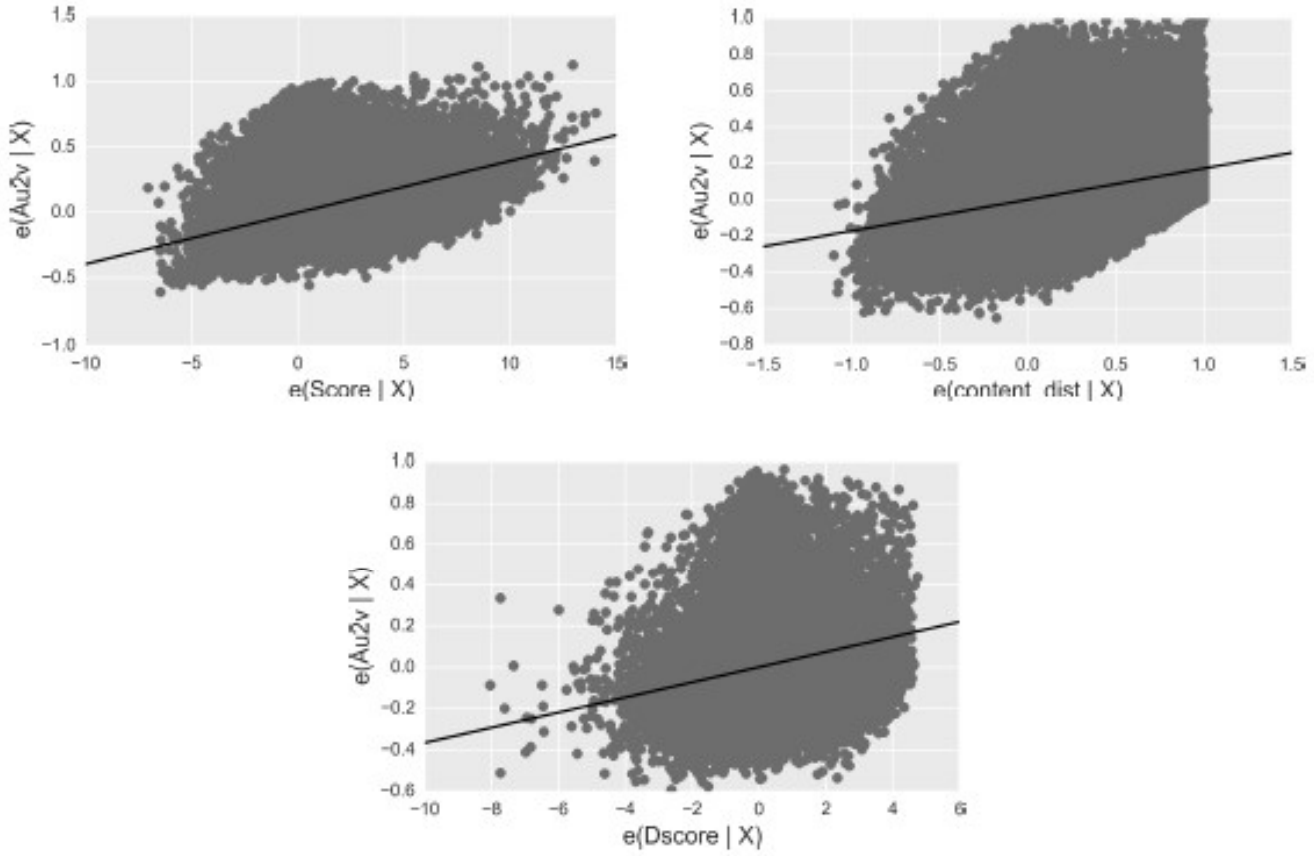
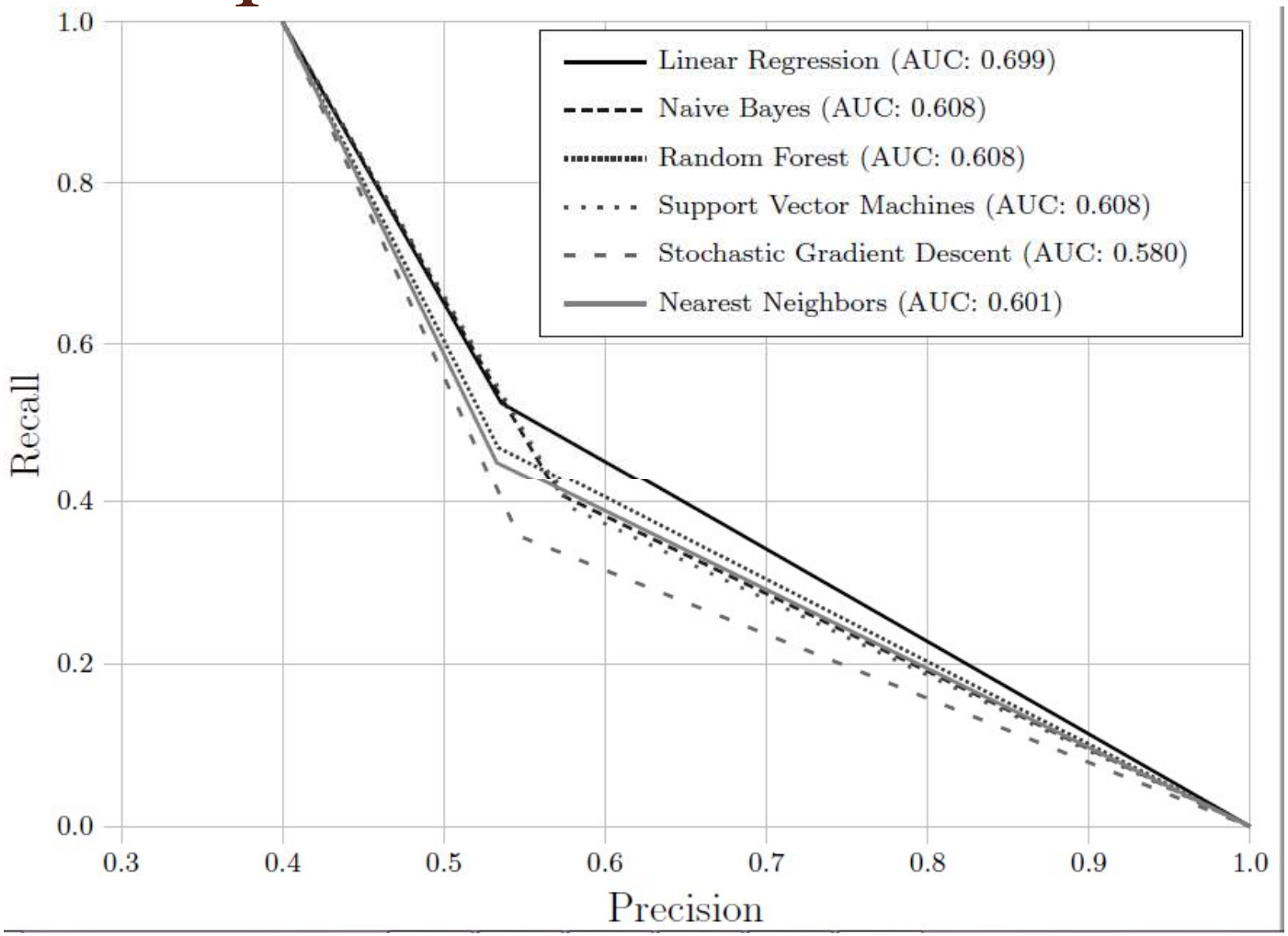


Figure 1. Regression plots for each independent variable.

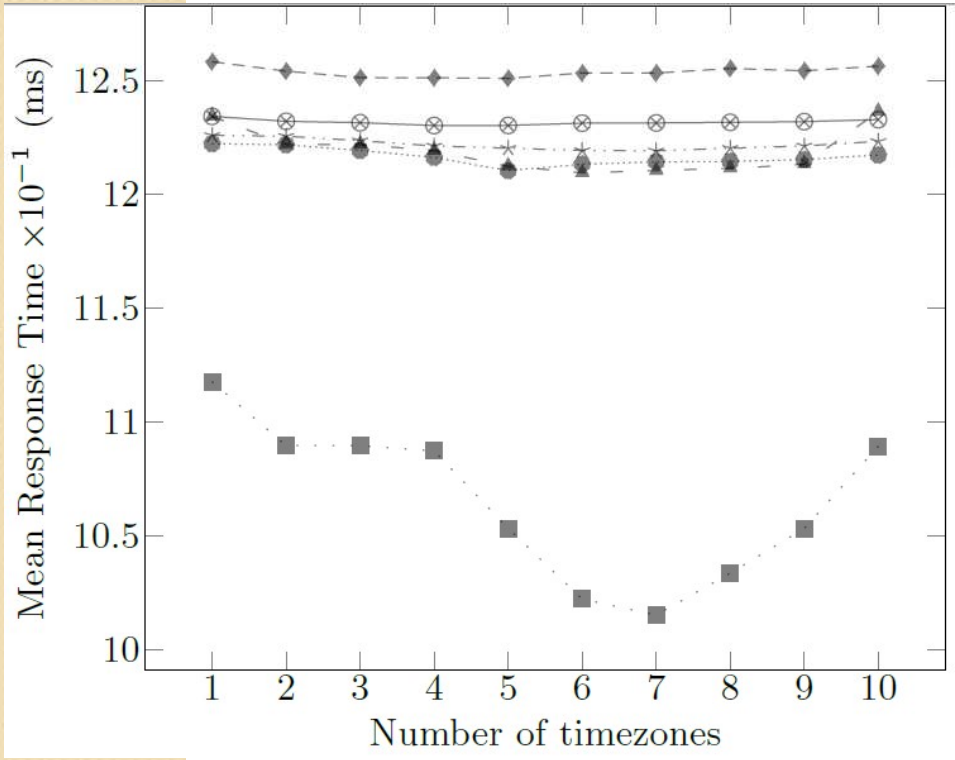
Regression results

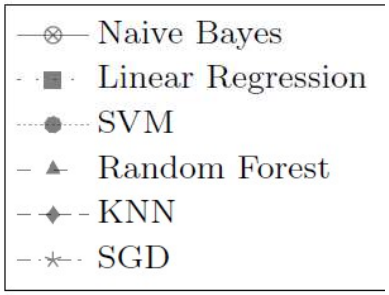
Comparison with other models



Experimental Evaluation –Main Findings

- Predictive Model: shift with approximately 7 zones out of the 10 used
- trade-off : MRT reduction- cost of copying in servers
  - cost for every copy related to the number of hops among the client and the server where copying is likely to take place
- Predictive Model: outperforms algorithms in [9], [10] (average MRT of 1.0647 msec)
- Precalculated zones with highest average values for each scheme





1: find the Y timezones that depict the highest average values of Predictive Model($Score$, $dScore$, $content\_dist$) user $V_i(t)$;

2: **for all** timezones that belong to Y **do**

3: find surrogate server n_j that serves timezone;

4: copy object o to n_j;

5: **end for**

Outline

INTRODUCTION AND MOTIVATION

- **Introductory Definitions**
- **Introduction and Motivation**

APPROACH

- **Phase 1: Literature Review on Content Delivery over OSNs**
- **Phase 2: The Social Prefetcher**
- **Phase 3: Parameterization**
- **Phase 4: Predictive Model for Diffusions over OSNs**

DISCUSSIONS AND CONCLUSIONS

- **Large-scale Datasets**
- **OSN Evolution**
- **Semantic Annotation**
- **Mobile CDNs and the Cloud**

Discussion and Conclusions

Mean response time	(Avg. 10⁻2 sec.)
Variation-1 - 24-h	1.1383
Variation-1 - 48-h	1.1352
Variation-1 - all-h	1.1172
LFU - all-h	1.1112
SIZE - all-h	1.1274
LRU - all-h	1.1172
Variation-2 - 24-h	1.1411
Variation-2 - 48-h	1.1376
Variation-2 - all-h	1.1174
Variation-3 (LR)- all-h	1.0647
Variation-3 (KNN)- all-h	1.2611
Variation-3 (NB)- all-h	1.2437
Variation-3 (SGD)- all-h	1.2364
Variation-3 (RF)- all-h	1.2252
Variation-3 (SVM)- all-h	1.2232
Social Prefetcher 24-h	1.1412
Social Prefetcher 48-h	1.1377
Social Prefetcher all-h	1.1181
Plain CDN Simulator - Push	1.4471
Plain CDN Simulator - Pull	1.8460

[175] I. Kilanioti, G.A. Papadopoulos, "Content Delivery Simulations supported by Social Network-awareness", Simulation Modelling Practice and Theory Journal - SIMPAT Elsevier, ChipSet Special Issue, under review

pull strategies achieve a lower mean response time under high loads, while push strategies are superior under low to medium load [171-173]

Discussion and Conclusions

- Conventional CDN systems vs Next generation CDN systems:
  - variety of social interactions with push and pull modes of information access
  - insufficient efforts to optimize multimedia CDN in the context of emerging social networks to enhance user experience
  - suboptimal on-demand high-quality video content delivery services – call for improved and additional CDN capabilities in place
  - advanced mechanisms for data placement, replication and distribution for a large variety of resource types and media formats
  - efficient handling of long-tail UGC (virality, localization)

-Large Scale Datasets

-OSN Evolution

-Semantic Annotation

-Mobile CDNs and the Cloud

  - discusses challenges inherent in developing OSN-aware content delivery applications
  - introduces novel algorithms for efficient delivery of UGC over OSNs
  - aims to serve as a starting point for extensive experimentation of the community with OSN-aware content delivery schemes

Thank you for your attention!

Questions?...

REFERENCES

[1] http://newsroom.fb.com/Key-Facts, [Online; accessed 17-April-2016].
[2] “Akamai. CDN services drive new Content Delivery Network (CDN) capabilities.” https://www.akamai.com/us/en/cdn.jsp, [Online; accessed 17-April-2016].
[3] “Center for Applied Internet Data Analysis,” https://www.caida.org, [Online; accessed 17-April-2016].
[4] “International Telecommunication Union. 2015. ict facts and figures. the world in 2015.” https://goo.gl/cqOSXA, [Online; accessed 17-April-2016].
[5] “Internet Society. Global Internet Report 2015. Mobile Evolution and Development of the Internet.” http://goo.gl/wUMJ8y, [Online; accessed 17-April-2016].
[6] “N. kennedy. facebooks photo storage rewrite.” http://www.niallkennedy.com/blog/2009/04/facebook-haystack.html.
[7] “The official Twitter blog.” https://blog.twitter.com/, [Online; accessed 17-April-2016].
[8] “Open source implementation of MapReduce.” http://hadoop.apache.org/, [Online; accessed 17-April-2016].
[9] “The top 500 sites on the web,” http://alexa.com/topsites, [Online; accessed 17-April-2016].
[10] “Twitter.” https://about.Twitter.com/company, [Online; accessed 17-April-2016].
[11] “Under the hood: Scheduling MapReduce jobs more efficiently with Corona, Facebook engineering,” https://goo.gl/ohfDe4, [Online; accessed 17-April-2016].
[12] “Vine.” [Online]. Available: https://vine.co
[13] “YouTube.” https://www.youtube.com/yt/about/, [Online; accessed 17-April-2016].
[14] “YouTube. Statistics.” https://www.youtube.com/yt/press/statistics.html, [Online; accessed 17-April-2016].
[15] A. S. A. Korn and A. Telcs, “Lobby index in networks,” Physica A: Statistical Mechanics and its Applications, vol. 388, no. 11, pp. 2221–2226, 2009.
[16] A. Abhari and M. Soraya, “Workload generation for YouTube,” Multimedia Tools Appl., vol. 46, no. 1, pp. 91–118, 2010. [Online]. Available: http://dx.doi.org/10.1007/s11042-009-0309-5
[17] A. Abisheva, V. R. K. Garimella, D. Garcia, and I. Weber, “Who watches (and shares) what on YouTube? and when?: using Twitter to understand Youtube viewership,” in Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014, 2014, pp. 593–602. [Online]. Available: http://doi.acm.org/10.1145/2556195.2566588
[18] C. Aggarwal, J. L. Wolf, and P. S. Yu, “Caching on the world wide web,” IEEE Transactions on Knowledge and Data Engineering, vol. 11, no. 1, pp. 94–107, Jan 1999.
[19] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, “Analysis of topological characteristics of huge online social networking services,” in Proceedings of the 16th international conference on World Wide Web. ACM, 2007, pp. 835–844.
[20] M. Aizenman and J. L. Lebowitz, “Metastability effects in Bootstrap Percolation,” Journal of Physics A: Mathematical and General, vol. 21, no. 19, p. 3801, 1999.
[21] N. Alon, M. Feldman, A. D. Procaccia, and M. Tennenholtz, “A note on competitive diffusion through social networks,” Information Processing Letters, vol. 110, no. 6, pp. 221–225, 2010.
[22] A. Anagnostopoulos, R. Kumar, and M. Mahdian, “Influence and correlation in social networks,” in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008, pp. 7–15.
[23] C. Anderson, Long Tail, The, Revised and Updated Edition: Why the Future of Business is Selling Less of More. Hyperion, 2008.
[24] W. B. Arthur, “Competing technologies, increasing returns, and lock-in by historical events,” The economic journal, vol. 99, no. 394, pp. 116–131, 1989.

[25] S. Asmussen, Applied probability and queues. Springer, 2003, vol. 51.

[26] B. Bahmani, R. Kumar, M. Mahdian, and E. Upfal, "PageRank on an evolving graph," in The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012, 2012, pp. 24–32. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339539

[27] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on Twitter," in Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011, 2011, pp. 65–74. [Online]. Available: http://doi.acm.org/10.1145/1935826.1935845

[28] E. Bakshy, I. Rosenn, C. Marlow, and L. A. Adamic, "The role of social networks in information diffusion," in Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012, 2012, pp. 519–528. [Online]. Available: http://doi.acm.org/10.1145/2187836.2187907

[29] R. Bandari, S. Asur, and B. Huberman, "The pulse of news in social media: Forecasting popularity," Arxiv preprint arXiv, vol. 1202, 2012.

[30] V. A. Banerjee, "A simple model of herd behavior," The Quarterly Journal of Economics, vol. 107, no. 3, pp. 797–817, 1992.

[31] A.-L. Barab´asi and R. Albert, "Emergence of scaling in random networks," science, vol. 286, no. 5439, pp. 509–512, 1999.

[32] F. M. Bass, "A new product growth for model consumer durables," Management Science, vol. 15, no. 5, pp. 215–227, 1969.

[33] G. Baxter, S. Dorogovtsev, A. Goltsev, and J. Mendes, "Bootstrap percolation on complex networks," Physical Review E, vol. 82, no. 1, p. 011103, 2010.

[34] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in Online Social Networks," in Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, 2009, pp. 49–62.

[35] E. Berger, "Dynamic monopolies of constant size," Journal of Combinatorial Theory, Series B, vol. 83, no. 2, pp. 191–200, 2001.

[36] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," Journal of political Economy, pp. 992–1026, 1992.

[37] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol. 1. IEEE, 1999, pp. 126–134.

[38] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," Computer networks and ISDN systems, vol. 30, no. 1, pp. 107–117, 1998.

[39] A. Brodersen, S. Scellato, and M. Wattenhofer, "YouTube around the world: geographic popularity of videos," in Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012, 2012, pp. 241–250. [Online]. Available: http://doi.acm.org/10.1145/2187836.2187870

[40] E. Brynjolfsson, Y. J. Hu, and M. Smith, "From niches to riches: Anatomy of the Long Tail," Sloan Management Review, vol. 47, no. 4, pp. 67–71, 2006.

[41] J. L. C. Huang, A. Wang and K. Ross, "Measuring and evaluating large-scale CDNs," in Internet Measurement Conference (IMC), 2008, 2008, pp. 15–29.

[42] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, 2007, pp. 1–14.

[43] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. B. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement 2007, San Diego, California, USA, October 24-26, 2007, 2007, pp. 1–14. [Online]. Available: http://doi.acm.org/10.1145/1298306.1298309

[44] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi, "Characterizing social cascades in Flickr," in Proceedings of the first workshop on Online Social Networks. ACM, 2008, pp. 13–18.

[45] K. Chard, S. Caton, O. Rana, and K. Bubendorfer, "Social Cloud: Cloud computing in social networks," in Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on. IEEE, 2010, pp. 99–106.

[46] J. Cheng, L. A. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, 2014, pp. 925–936. [Online]. Available: http://doi.acm.org/10.1145/2566486.2567997

[47] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of YouTube videos," in 16th International Workshop on Quality of Service, IWQoS 2008, University of Twente, Enskede, The Netherlands, 2-4 June 2008., 2008, pp. 229–238. [Online]. Available: http://dx.doi.org/10.1109/IWQOS.2008.32

[48] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings algorithm," The American Statistician, vol. 49, no. 4, pp. 327–335, 1995.

[49] F. Chierichetti, J. Kleinberg, and A. Panconesi, "How to schedule a cascade in an arbitrary graph," in Proceedings of the 13th ACM Conference on Electronic Commerce. ACM, 2012, pp. 355–368.

[50] P. Clifford and A. Sudbury, "A model for spatial conflict," Biometrika, vol. 60, no. 3, pp. 581–588, 1973.

[51] D. Daley and D. G. Kendall, “Stochastic rumours,” IMA Journal of Applied Mathematics, vol. 1, no. 1, pp. 42–55, 1965.

[52] K. S. Dave, R. Bhatt, and V. Varma, “Modelling action cascades in social networks,” 5th ICWSM, 2011.

[53] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.

[54] L. Dickens, I. Molloy, J. Lobo, P.-C. Cheng, and A. Russo, “Learning stochastic models of information flow,” in Data Engineering (ICDE), 2012 IEEE 28th International Conference on. IEEE, 2012, pp. 570–581.

[55] Y. Dodge, D. Cox, D. Commenges, A. Davison, P. Solomon, and S. Wilson, The Oxford dictionary of statistical terms. Oxford University Press, USA, 2006.

[56] C. Doerr, S. Tang, N. Blenn, and P. Van Mieghem, “Are friends overrated,” A Study of the Social News Aggregator Digg. com (IFIP Networking, 2011), 2011.

[57] P. Domingos and M. Richardson, “Mining the network value of customers,” in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001, pp. 57–66.

[58] P. A. Dow, L. A. Adamic, and A. Friggeri, “The anatomy of large Facebook cascades,” in Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013., 2013. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6123

[59] M. Draief, A. Ganesh, and L. Massouli´e, “Thresholds for virus spread on networks,” in Proceedings of the 1st international conference on Performance evaluation methodolgies and tools. ACM, 2006, p. 51.

[60] D. Easley and J. Kleinberg, Networks, crowds, and markets. Cambridge Univ Press, 2010.

[61] D. A. Easley and J. M. Kleinberg, Networks, Crowds, and Markets - Reasoning About a Highly Connected World. Cambridge University Press, 2010. [Online]. Available: http://www.cambridge.org/gb/knowledge/isbn/item2705443/?site locale=en GB

[62] P. ERDdS and A. R&WI, “On random graphs i.” Publ. Math. Debrecen, vol. 6, pp. 290–297, 1959.

[63] E. Even-Dar and A. Shapira, “A note on maximizing the spread of influence in social networks,” Internet and Network Economics, pp. 281–286, 2007.

[64] A. Fard, A. Abdolrashidi, L. Ramaswamy, and J. A. Miller, “Towards efficient query processing on massive time-evolving graphs,” in Proceedings of the 8th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), Pittsburgh, PA, United States, 2012.

[65] F. Figueiredo, F. Benevenuto, and J. M. Almeida, “The tube over time: characterizing popularity growth of YouTube videos,” in Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011, 2011, pp. 745–754. [Online]. Available: http://doi.acm.org/10.1145/1935826.1935925

[66] A. Finamore, M. Mellia, M. M. Munafˆo, R. Torres, and S. G. Rao, “YouTube everywhere: impact of device and infrastructure synergies on user experience,” in Proceedings of the 11th ACM SIGCOMM Conference on Internet Measurement, IMC ’11, Berlin, Germany, November 2-, 2011, 2011, pp. 345–360. [Online]. Available: http://doi.acm.org/10.1145/2068816.2068849

[67] J. Fowler and N. Christakis, “Connected: The surprising power of our social networks and how they shape our lives,” HarperCollins Publishers, 2009.

[68] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, “Outtweeting the Twitterers - Predicting Information Cascades in Microblogs,” in 3rd Workshop on Online Social Networks, WOSN 2010, Boston, MA, USA, June 22, 2010, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1863190.1863193

[69] A. Garcia-Silva, J.-H. Kang, K. Lerman, and O. Corcho, “Characterising emergent semantics in Twitter lists,” in The Semantic Web: Research and Applications. Springer, 2012, pp. 530–544.

[70] E. N. Gilbert, “Random graphs,” The Annals of Mathematical Statistics, pp. 1141–1144, 1959.

[71] P. Gill, M. F. Arlitt, Z. Li, and A. Mahanti, “YouTube traffic characterization: a view from the edge,” in Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement 2007, San Diego, California, USA, October 24-26, 2007, 2007, pp. 15–28. [Online]. Available: http://doi.acm.org/10.1145/1298306.1298310

[72] J. Goldenberg, B. Libai, and E. Muller, “Talk of the network: A complex systems look at the underlying process of word-of-mouth,” Marketing letters, vol. 12, no. 3, pp. 211–223, 2001.

[73] A. Goyal, F. Bonchi, and L. V. Lakshmanan, “Learning influence probabilities in social networks,” in Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010, pp. 241–250.

[74] M. Granovetter, “Threshold models of collective behavior,” American journal of sociology, pp. 1420–1443, 1978.

[75] P. Grindrod, M. C. Parsons, D. J. Higham, and E. Estrada, “Communicability across evolving networks,” Physical Review E, vol. 83, no. 4, p. 046120, 2011.

[76] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, “Information diffusion through blogspace,” in Proceedings of the 13th international conference on World Wide Web. ACM, 2004, pp. 491–501.

[77] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, “WTF: the who to follow service at Twitter,” in 22nd International World Wide Web Conference, WWW ’13, Rio de Janeiro, Brazil, May 13-17, 2013, 2013, pp. 505–514. [Online]. Available: http://dl.acm.org/citation.cfm?id=2488433

[78] T. Hogg and K. Lerman, “Stochastic models of user-contributory web sites,” in Proc. Int. Conference on Weblogs and Social Media, 2009, pp. 95–97.

[79] R. A. Holley and T. M. Liggett, “Ergodic theorems for weakly interacting infinite systems and the voter model,” The annals of probability, pp. 643–663, 1975.

[80] L. Hong, O. Dan, and B. D. Davison, “Predicting popular messages in Twitter,” in Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011 (Companion Volume), 2011, pp. 57–58. [Online]. Available: http://doi.acm.org/10.1145/1963192.1963222

[81] I. Hoque and I. Gupta, “Disk layout techniques for online social network data,” Internet Computing, IEEE, vol. 16, no. 3, pp. 24–36, 2012.

[82] K. Huguenin, A.-M. Kermarrec, K. Kloudas, and F. Ta¨ıani, “Content and geographical locality in user-generated content sharing systems,” in Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video. ACM, 2012, pp. 77–82.

[83] E. Ising, “Beitrag zur Theorie des Ferromagnetismus,” Zeitschrift f¨ur Physik A Hadrons and Nuclei, vol. 31, no. 1, pp. 253–258, 1925.

[84] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. Briggs, and R. Braynard, “Networking named content,” Commun. ACM, vol. 55, no. 1, pp. 117–124, 2012. [Online]. Available: http://doi.acm.org/10.1145/2063176.2063204

[85] A. Java, X. Song, T. Finin, and B. Tseng, “Why we Twitter: understanding microblogging usage and communities,” in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, 2007, pp. 56–65.

[86] M. Jenders, G. Kasneci, and F. Naumann, “Analyzing and predicting viral tweets,” in 22nd International World Wide Web Conference, WWW ’13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume, 2013, pp. 657–664. [Online]. Available:

[87] U. Kang, H. Tong, J. Sun, C.-Y. Lin, and C. Faloutsos, “Gbase: a scalable and general graph management system,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2011, pp. 1091–1099.

[88] U. Kang, C. E. Tsourakakis, and C. Faloutsos, “Pegasus: A peta-scale graph mining system implementation and observations,” in Data Mining, 2009. ICDM’09. Ninth IEEE International Conference on. IEEE, 2009, pp. 229–238.

[89] E. Katz and P. F. Lazarsfeld, Personal influence: The part played by people in the flow of mass communications. Transaction Pub, 2006.

[90] D. Kempe, J. Kleinberg, and ´ E. Tardos, “Maximizing the spread of influence through a social network,” in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003, pp. 137–146.

[91] D. Kempe, J. Kleinberg, and E. Tardos, “Influential nodes in a diffusion model for social networks,” Automata, Languages and Programming, pp. 99–99, 2005.

[92] I. Kilanioti, “Improving multimedia content delivery via augmentation with social information. The Social Prefetcher approach.” Multimedia, IEEE Transactions on, vol. 17, no. 9, pp. 1460–1470, 2015. [Online]. Available: http://goo.gl/x81Xv1

[93] I. Kilanioti and G. A. Papadopoulos, “Socially-aware multimedia content delivery for the cloud,” in 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC), Dec 2015, pp. 300–309. [Online]. Available: http://goo.gl/gsjh0E

[94] I. Kilanioti, C. Georgiou, and G. Pallis, “On the impact of online social networks in content delivery,” in Advanced Content Delivery and Streaming in the Cloud, M. Pathan, R. Sitaraman, and D. Robinson, Eds. Wiley, 2014.

[95] J. Kleinberg, “Navigation in a Small World,” Nature, vol. 406, no. 6798, pp. 845–845, 2000.

[96] J. M. Kleinberg, “Authoritative sources in an hyperlinked environment,” J. ACM, vol. 46, no. 5, pp. 604–632, 1999. [Online]. Available: http://doi.acm.org/10.1145/324133.324140

[97] J. Kleinberg, “Cascading behavior in networks: Algorithmic and economic issues,” in Algorithmic game theory (N. Nisan, T. Roughgarden, E. Tardos, V. Vazirani, eds.). Cambridge University Press, 2007, pp. 613–632.

[98] R. Kumar, J. Novak, and A. Tomkins, “Structure and evolution of Online Social Networks,” Link Mining: Models, Algorithms, and Applications, pp. 337–357, 2010.

[99] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev, "Prediction of retweet cascade size over time," in 21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012, 2012, pp. 2335–2338. [Online]. Available: http://doi.acm.org/10.1145/2396761.2398634

[100] M. Kuperman and G. Abramson, "Small world effect in an epidemiological model," Physical Review Letters, vol. 86, no. 13, pp. 2909–2912, 2001.

[101] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 591–600.

[102] H. Kwak, C. Lee, H. Park, and S. B. Moon, "What is Twitter, a social network or a news media?" in Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010, 2010, pp. 591–600. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772751

[103] T. La Fond and J. Neville, "Randomization tests for distinguishing social influence and homophily effects," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 601–610.

[104] K. Lerman, S. Intagorn, J.-H. Kang, and R. Ghosh, "Using proximity to predict activity in social networks," in Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012, pp. 555–556.

[105] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in Proceeding of the 17th international conference on World Wide Web. ACM, 2008, pp. 915–924.

[106] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 1, p. 2, 2007.

[107] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Patterns of cascading behavior in large blog graphs," in Proceedings of SIAM International Conference on Data Mining (SDM) 2007. SIAM, 2007.

[108] Y. Li, Y. Shen, and Y. Liu, "Utilizing content delivery network in cloud computing," in Computational Problem-Solving (ICCP), 2012 International Conference on, Oct 2012, pp. 137–143.

[109] D. Liben-Nowell and J. Kleinberg, "Tracing information flow on a global scale using Internet chain-letter data," Proceedings of the National Academy of Sciences, vol. 105, no. 12, pp. 4633–4638, 2008.

[110] C. X. Lin, Q. Mei, Y. Jiang, J. Han, and S. Qi, "Inferring the diffusion and evolution of topics in social communities," mind, vol. 3, no. d4, p. d5, 2011.

[111] I. Lobel, M. Dahleh, D. Acemoglu, and A. Ozdaglar, "Bayesian learning in social networks," NYU Working Paper No. CEDER-09-01, 2009.

[112] T. Łuczak, "Size and connectivity of the K-core of a random graph," Discrete Mathematics, vol. 91, no. 1, pp. 61–68, 1991.

[113] Z. Ma, A. Sun, and G. Cong, "On predicting the popularity of newly emerging hashtags in Twitter," JASIST, vol. 64, no. 7, pp.

1399–1410, 2013. [Online]. Available: http://dx.doi.org/10.1002/asi.22844

[114] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," Annual review of sociology, pp. 415–444, 2001.

[115] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in Proceeding of the 17th international

conference on World Wide Web, 2008, pp. 101–110.

[116] S. Milgram, "The Small World problem," Psychology today, vol. 2, no. 1, pp. 60–67, 1967.

[117] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks,"

in Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, 2007, pp. 29–42.

[118] S. Mitra, M. Agrawal, A. Yadav, N. Carlsson, D. L. Eager, and A. Mahanti, "Characterizing web-based video sharing workloads," TWEB, vol. 5, no. 2, pp. 1–27, 2011. [Online]. Available: http://doi.acm.org/10.1145/1961659.1961662

[119] J. Mondal and A. Deshpande, "Managing large dynamic graphs efficiently," in Proceedings of the 2012 international conference on Management of Data. ACM, 2012, pp. 145–156.

[120] S. Morris, "Contagion," The Review of Economic Studies, vol. 67, no. 1, pp. 57–78, 2000.

[121] E. Mossel and S. Roch, "On the submodularity of influence in social networks," in Proceedings of the thirty-ninth annual ACM symposium on Theory of computing. ACM, 2007, pp. 128–134.

[122] S. A. Myers and J. Leskovec, "The bursty dynamics of the Twitter information network," in 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, 2014, pp. 913–924. [Online]. Available: http://doi.acm.org/10.1145/2566486.2568043

[123] A. Najar, L. Denoyer, and P. Gallinari, "Predicting information diffusion on social networks with partial knowledge," in Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012, pp. 1197–1204.

[124] R. Narayanam and Y. Narahari, "A shapley value-based approach to discover influential nodes in social networks," IEEE Transactions on Automation Science and Engineering, no. 99, pp. 1–18, 2010.

[125] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, "Theory of rumour spreading in complex social networks," Physica A: Statistical Mechanics and its Applications, vol. 374, no. 1, pp. 457–470, 2007.

[126] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functionsI," Mathematical Programming, vol. 14, no. 1, pp. 265–294, 1978.

[127] Z. L. P. Gill, M. Arlitt and A. Mahanti, "Characterizing user sessions on YouTube," in ACM/SPIE Multimedia Computing and Networking Conference (MMCN '08), San Jose, USA, 2008, 2008.

[128] G. Peng, "CDN: Content distribution network technical report tr-125," Experimental Computer Systems Lab, Stony Brook University, 2003.

[129] S. Petrovic, M. Osborne, and V. Lavrenko, "RT to win! predicting message propagation in Twitter," in Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011, 2011. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2754

[130] A. Rao, A. Legout, Y. Lim, D. Towsley, C. Barakat, and W. Dabbous, "Network characteristics of video streaming traffic,"

in Proceedings of the 2011 Conference on Emerging Networking Experiments and Technologies, Co-NEXT '11, Tokyo, Japan, December 6-9, 2011, 2011, p. 25. [Online]. Available: http://doi.acm.org/10.1145/2079296.2079321

[131] C. Ren, E. Lo, B. Kao, X. Zhu, and R. Cheng, "On querying historial evolving graph sequences," Proceedings of the VLDB Endowment, vol. 4, no. 11, 2011.

[132] T. Rodrigues, F. Benevenuto, M. Cha, P. K. Gummadi, and V. A. F. Almeida, "On word-of-mouth based discovery of the web," in Proceedings of the 11th ACM SIGCOMM Conference on Internet Measurement, IMC '11, Berlin, Germany, November 2-, 2011, 2011, pp. 381–396. [Online]. Available: http://doi.acm.org/10.1145/2068816.2068852

[133] M. G. Rodriguez, J. Leskovec, and B. Sch¨olkopf, "Structure and dynamics of information pathways in online media," in Proceedings of ACM International Conference on Web Search and Data Mining (WSDM), Rome, Italy, 2013.

[134] E. M. Rogers, Diffusion of innovations. Simon and Schuster, 1995.

[135] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Learning continuous-time information diffusion model for social behavioral data analysis," Advances in Machine Learning, pp. 322–337, 2009.

[136] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for Independent Cascade Model," in Knowledge- Based Intelligent Information and Engineering Systems. Springer, 2008, pp. 67–75.

[137] N. Sastry, E. Yoneki, and J. Crowcroft, "Buzztraq: predicting geographical access patterns of social cascades using social networks," in Proceedings of the Second ACM EuroSys Workshop on Social Network Systems. ACM, 2009, pp. 39–45.

[138] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft, "Track globally, deliver locally: improving Content Delivery Networks by tracking geographic social cascades," in Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011, 2011, pp. 457–466. [Online]. Available: http://doi.acm.org/10.1145/1963405.1963471

[139] D. Schi¨oberg, F. Schneider, G. Tredan, S. Uhlig, and A. Feldmann, "Revisiting content availability in Distributed Online SocialNetworks."

[140] T. C. Sendling, "Micromotives and macrobehavior," New York and London: Norton, 1978.

[141] X. Song, Y. Chi, K. Hino, and B. L. Tseng, "Information flow modeling based on diffusion rate for prediction and ranking," in Proceedings of the 16th international conference on World Wide Web. ACM, 2007, pp. 191–200.

[142] F. Sorrentino, M. Di Bernardo, F. Garofalo, and G. Chen, "Pinning-controllability of complex networks," arXiv preprint condmat/0701073, 2007.

[143] K. Stamos, G. Pallis, A. Vakali, D. Katsaros, A. Sidiropoulos, and Y. Manolopoulos, "CDNsim: A simulation tool for Content Distribution Networks," ACM Trans. Model. Comput. Simul., vol. 20, no. 2, pp. 1–40, 2010. [Online]. Available: http://doi.acm.org/10.1145/1734222.1734226

[144] D. Stauffer and A. Aharony, Introduction to Percolation Theory. CRC, 1994.

.

[145] G. V. Steeg, R. Ghosh, and K. Lerman, "What stops social epidemics?" arXiv preprint arXiv:1102.1985, 2011.

[146] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network," in Proceedings of the 2010 IEEE Second International Conference on Social Computing, SocialCom / IEEE International Conference on Privacy, Security, Risk and Trust, PASSAT 2010, Minneapolis, Minnesota, USA, August 20-22, 2010, 2010, pp. 177–184. [Online]. Available: http://dx.doi.org/10.1109/SocialCom.2010.33

[147] G. Szab´o and B. A. Huberman, "Predicting the popularity of online content," Commun. ACM, vol. 53, no. 8, pp. 80–88, 2010. [Online]. Available: http://doi.acm.org/10.1145/1787234.1787254

[148] W. Tang, Y. Fu, L. Cherkasova, and A. Vahdat, "MediSyn: a synthetic streaming media service workload generator," in Network and Operating System Support for Digital Audio and Video, 13th International Workshop, NOSSDAV 2003, Monterey, CA, USA, June 1-3, 2003, Proceedings, 2003, pp. 12–21. [Online]. Available: http://doi.acm.org/10.1145/776322.776327

[149] R. Torres, A. Finamore, J. R. Kim, M. Mellia, M. M. Munaf`o, and S. G. Rao, "Dissecting video server selection strategies in the YouTube CDN," in 2011 International Conference on Distributed Computing Systems, ICDCS 2011, Minneapolis, Minnesota, USA, June 20-24, 2011, 2011, pp. 248–257. [Online]. Available: http://dx.doi.org/10.1109/ICDCS.2011.43

[150] D. Towsley, A. Rao, Y.-S. Lim, C. Barakat, A. Legout, and W. Dabbous, "Network characteristics of video streaming traffic," in Proceedings of the 7th Conference on Emerging Networking Experiments and Technologies (CoNEXT), New York, NY, USA, 2011.

[151] S. Traverso, K. Huguenin, I. Trestian, V. Erramilli, N. Laoutaris, and K. Papagiannaki, "TailGate: handling long-tail content with a little help from friends," in Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012, 2012, pp. 151–160. [Online]. Available: http://doi.acm.org/10.1145/2187836.2187858

[152] S. Triukose, Z. Wen, and M. Rabinovich, "Measuring a commercial Content Delivery Network," in Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011, 2011, pp. 467–476. [Online]. Available: http://doi.acm.org/10.1145/1963405.1963472

[153] O. Tsur and A. Rappoport, "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities," in Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012, 2012, pp. 643–652. [Online]. Available: http://doi.acm.org/10.1145/2124295.2124320

[154] C. Van den Bulte and Y. V. Joshi, "New product diffusion with influentials and imitators," Marketing Science, vol. 26, no. 3, pp. 400–421, 2007.

[155] S. Wasserman and K. Faust, Social Network Analysis: Methods and applications. Cambridge university press, 1994, vol. 8.

[156] D. Watts and S. Strogatz, "The Small World problem," Collective Dynamics of Small-World Networks, vol. 393, pp. 440–442, 1998.

[157] D. J. Watts, "A simple model of global cascades on random networks," Proceedings of the National Academy of Sciences, vol. 99, no. 9, pp. 5766–5771, 2002.

[158] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential Twitterers," in Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010, pp. 261–270.

[159] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in Proceedings of the 4th ACM European conference on Computer systems. Acm, 2009, pp. 205–218

[160] F. Wu and B. A. Huberman, “Novelty and collective attention,” Proceedings of the National Academy of Sciences, vol. 104, no. 45, pp. 17 599–17 601, 2007.

[161] J. Yang and J. Leskovec, “Modeling information diffusion in implicit networks,” in Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010, pp. 599–608.

[162] S. Yang and H. Zha, “Mixture of mutually exciting processes for viral diffusion,” in Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013, 2013, pp. 1–9. [Online]. Available: http://jmlr.org/proceedings/papers/v28/yang13a.html

[163] J. Yang and J. Leskovec, “Patterns of temporal variation in online media,” in Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011, 2011, pp. 177–186. [Online]. Available: http://doi.acm.org/10.1145/1935826.1935863

[164] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, “Understanding user behavior in large-scale video-on-demand systems,” in ACM SIGOPS Operating Systems Review, vol. 40, no. 4. ACM, 2006, pp. 333–344.

[165] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern, “Predicting information spreading inTwitter,” in Workshop on computational social science and the wisdom of crowds, nips, vol. 104, no. 45. Citeseer, 2010, pp. 17 599–601.

[166] F. Zhou, L. Zhang, E. Franco, A. Mislove, R. Revis, and R. Sundaram, “WebCloud: Recruiting social network users to assist in content distribution,” in Proceedings of IEEE International Symposium on Network Computing and Applications, Cambridge, MA, USA, 2012.

[167] I. Kilanioti and G. A. Papadopoulos, “Efficient content delivery through popularity forecasting on social media,” in Proceedings of the 7th IEEE International Conference on Information, Intelligence, Systems and Applications, IISA 2016, Chalkidiki, Greece, July 13-15, 2016, pp. 13–19.

[168] I. Kilanioti and G. A. Papadopoulos, “Delivering social multimedia content with scalability,” in Resource Management for Big Data Platforms: Algorithms, Modelling and High-Performance Computing Techniques, Springer Computer Communications and Networks Series, Eds. F. Pop, J. Kolodjiez, B. D. Martino, Springer, 2016.

[169]] I. Kilanioti and G. A. Papadopoulos, “Predicting video virality on Twitter,” in Resource Management for Big Data Platforms: Algorithms, Modelling and High-Performance ComputingTechniques, Springer Computer Communications and Networks Series, Eds. F. Pop, J. Kolodjiez, B. D. Martino, Springer, 2016.

[170] T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau, “Internet video delivery in youtube: From traffic measurements to quality of experience,” in Data Traffic Monitoring and Analysis. sSpringer, 2013, pp. 264–301.

[171] D. L. Eager, E. D. Lazowska, and J. Zahorjan, “A comparison of receiver-initiated and sender-initiated adaptive load sharing,” Perform. Eval., vol. 6, no. 1, pp. 53–68, 1986.

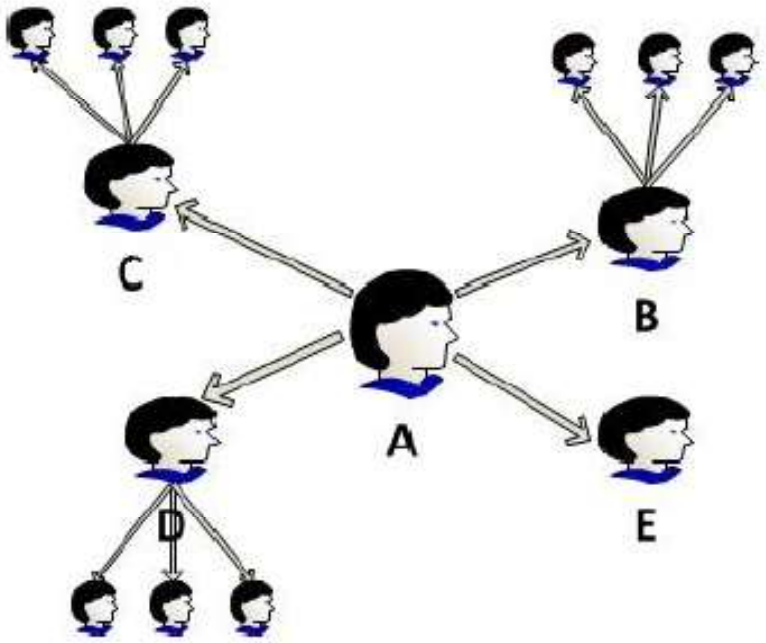
[172] R. Mirchandaney, D. Towsley, and J. A. Stankovic, “Adaptive load sharing in heterogeneous distributed systems,” J. Parallel Distrib. Comput., vol. 9, no. 4, pp. 331–346, 1990.

[173] W. Minnebo and B. Van Houdt, “Pull versus push mechanism in large distributed networks: Closed form results” in Proceedings of the 24th International Teletraffic Congress, ser. ITC ’12. International Teletraffic Congress, 2012, pp. 9:1–9:8.

[174] J. Cheng, L. A. Adamic, J. M. Kleinberg, and J. Leskovec, “Do cascades recur?”in Proceedings of the 25th International Conference on World Wide Web, ser. WWW ’16. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016, pp. 671–681.

[175] I. Kilanioti, G.A. Papadopoulos, “Content Delivery Simulations supported by Social Network-awareness”, Simulation Modelling Practice and Theory Journal - SIMPAT Elsevier, ChipSet Special Issue, under review

Lobby index example – HITS example



Largest integer l such that v has l neighbours with a degree of at least l.

l(A) =3

$$hub\_score_A = 0.999985,$$
$$authority\_score_A = 0,\ hub\_score_B = 0.00317,\ authority\_score_B = 0.499990,$$
$$hub\_score_C = 0.003171,\ authority\_score_C = 0.499990,\ hub\_score_D = 0.003171,$$
$$authority\_score_D = 0.499990,\ hub\_score_E = 0,\ authority\_score_E = 0.499990.$$