Weighted Overlay Design for Topic-based Pub/Sub on Geo-Distributed Data Centers

Chen Chen
IBM Research – Haifa
chenc@il.ibm.com

Joint work with Yoav Tock, IBM Research – Haifa
Hans-Arno Jacobsen, University of Toronto
Roman Vitenberg, University of Oslo
Publish/Subscribe (pub/sub)

Publish:
topic = IBM
price = $170

Publish:
topic = Microsoft
price = $50

Subscribe:
IBM

Subscribe:
Microsoft
Pub/Sub

• Many applications and standards
  – Online social networks, financial data dissemination, application integration, RSS feed distribution, business process management
  – MQTT, XMPP, etc.

• Industrial adoptions
  – IBM Internet of Things (IoT) Foundation
  – Google GooPS
  – Yahoo PNUTS
  – Facebook
  – Twitter
Two major directions to improve pub/sub

Overlay topologies
The construction of the overlay topology such that network traffic is minimized
• Chockler et al., Podc’07
• Onus et al., Infocom’09

Routing protocols
The design of routing protocols so that messages are disseminated efficiently on the overlay network
• G. Li et al., Icdcs’08
• M. Castro et al., J sac’02
Two major directions to improve pub/sub

Overlay topologies

Routing protocols
Two major directions to improve pub/sub

Overlay topologies

Routing protocols
Topic-connected overlay (TCO)

An overlay $G$
Topic-connected overlay (TCO)

An overlay $G$ is topic-connected with one *TC-component* $G^{(a)}$. 

$G^{(a)}$ is topic-connected with one *TC-component*.
An overlay $G$ is topic-connected with one $TC$-component $G^{(a)}$ is topic-connected with two $TC$-components $G^{(b)}$ is NOT topic-connected
Why TCO?

• Enables efficient pub/sub routing
  – Zero relay on non-interested nodes
  – Simpler matching engine
  – Smaller forwarding tables

• Supports advanced security
  – users in a trusted group can share sensitive data among themselves without having it travel outside the group
Drawbacks of TCO

• Implicit assumption of edge equivalence
  – oblivious to underlying network infrastructure
  – regardless of locality
• Not applicable for IBM IoT Foundation
  – O(100M) IoT devices
  – geographically distributed data centers
• We need a better pub/sub overlay design
Our contributions

• Extend TCO to Weighted TCO (WTCO) locality-aware pub/sub overlays
• Propose MinAvg-WTCO problem
  – NP-completeness
  – lower bound for approximation
• Design approximation algorithms
  – greedy, primal-dual, etc.
  – divide-and-conquer
• Conduct empirical evaluation
Weighted TCO (WTCO) model

• Topic-connected overlay (TCO)
  – $V$, a set of nodes
  – $T$, a set of topics
  – $I$, an interest matrix
  – $E$, a set of overlay edges

• Weighted TCO (WTCO)
  – $\omega$, an edge weight function
Weighted TCO (WTCO) model

• Topic-connected overlay (TCO)
  – \( V \), a set of nodes
  – \( T \), a set of topics
  – \( I \), an interest matrix
  – \( E \), a set of overlay edges

• Weighted TCO (WTCO)
  – \( \omega \), an edge weight function

- network latencies
- bandwidths
- costs of traffic
MinAvg-WTCO problem

• Input \((V, T, I, \omega)\)
• Output \(E \subseteq K\), where \(K = V \times V\)

An edge set that forms a WTCO for \((V, T, I, \omega)\) with the smallest weighted average node degree, i.e., the sum of all edge weights is minimum

\[
\min_{E \subseteq K} \left\{ \sum_{e \in E} \omega(e) : E \text{ forms a WTCO for } (V, T, I, \omega) \right\}
\]
Greedy algorithm for MinAvg-WTCO

GrMAW

Input: \((V, T, I, \omega)\)

Output: \(E\)

1. \(E \leftarrow \emptyset\)

2. while \(E\) does not form a WTCO for \((V, T, I, \omega)\)

3. greedily add to \(E\) the most cost-effective edge

4. return \(E\)
Greedy algorithm for MinAvg-WTCO

GrMAW

Input: \((V, T, I, \omega)\)
Output: \(E\)

1. \(E \leftarrow \emptyset\)
2. while \(E\) does not form a WTCO for \((V, T, I, \omega)\)
3. greedily add to \(E\) the most cost-effective edge
4. return \(E\)
“Greed is good”
“Greed is good”

• **GrMAW**: polynomial running time and logarithmic approximation ratio

\[ O(\log |T|) \]
“Greed is good”

• **GrMAW**: polynomial running time and logarithmic approximation ratio

\[ O(\log |T|) \]

• **MinAvg-WTCO** is NP-complete. If P ≠ NP, the approximation ratio of any polynomial-time algorithm for MinAvg-WTCO is

\[ \Omega(\log |V|) \]
Proof of approximation in a nutshell

• Cast MinAvg-WTCO into the form of Submodular Set Cover

• Analyze the greedy algorithm GrMAW based on an existing theorem
Progress measure of GrMAW
Progress measure of GrMAW

$TCC(E)$: # topic-connected components (TCCs) over all topics w.r.t. the edge set $E$

$$TCC(E) = \sum_{t \in T} \left( \# \text{TCCs in } G^{(t)} = (V^{(t)}, E^{(t)}) \right)$$
Progress measure of GrMAW

\[ TCC(E) \]: \# topic-connected components (TCCs) over all topics w.r.t. the edge set \( E \)

\[
TCC(E) = \sum_{t \in T} \left( \# \text{TCCs in } G^{(t)} = (V^{(t)}, E^{(t)}) \right)
\]
Contribution of edge

• Contribution (towards WTCO) of edge $e$: 
  
  $\#TCCs$ that would be reduced by adding $e$ to $E$

  \[
  contrib_E(e) = TCC(E) - TCC(E + e)
  \]
Contribution of edge

- Contribution (towards WTCo) of edge $e$: 
  
  \#TCCs that would be reduced by adding $e$ to $E$

  \[
  \text{contrib}_E(e) = TCC(E) - TCC(E + e)
  \]
\[ \mu(E) = TCC(\emptyset) - TCC(E) \]

Submodular

\[ \mu(E + e) - \mu(E) \geq \mu(F + e) - \mu(F), \forall E \subseteq F, \forall e \]
\[
\mu(E) = TCC(\emptyset) - TCC(E)
\]

Submodular

\[
\mu(E + e) - \mu(E) \geq \mu(F + e) - \mu(F), \forall E \subseteq F, \forall e
\]

\[\text{contrib}_E(e) \quad \text{contrib}_F(e)\]
$$\mu(E) = TCC(\emptyset) - TCC(E)$$

**Submodular**

$$\mu(E + e) - \mu(E) \geq \mu(F + e) - \mu(F), \forall E \subseteq F, \forall e$$

\text{contribution} \quad \mu(E + e) - \mu(E) \geq \mu(F + e) - \mu(F), \forall E \subseteq F, \forall e$

**Diminishing return**: contribution of $e$ decreases, as current edge set progresses from $E$ to $F$
MinAvg-WTCO --> SSC

MinAvg-WTCO

$$\min_{E \subseteq K} \left\{ \sum_{e \in E} \omega(e) : E \text{ forms a WTCO for } (V, T, I, \omega) \right\}$$

Submodular Set Cover (SSC)

$$\min_{E \subseteq K} \left\{ \sum_{e \in E} \omega(e) : \mu(E) = \mu(K) \right\}$$
Approximation Ratio of GrMAW

GrMAW achieves the approximation ratio

\[ O(\log |T|) \]

Theorem: greediness attains a logarithmic approximation ratio for SSC
Evaluation

• Number of nodes $|V|$: [1000, 10000]
• Number of topic $|T|$: [100, 1000]
• Subscription size on average: [50, 150]
• Topic popularity
  - Uniform: [Chockler et al., 2007]
  - Zipf: news feed in RSS [Liu et al., 2005]
  - Exponential: stock in NYSE [Tock et al., 2005]
• Edge weight function:
  - King [Gummadi et al., 2002]
  - Meridian [Wong et al., 2005]
Greedy versus Tree-per-topic

Weighted average degree vs. #nodes graph comparing Greedy (GrMAW) and Tree-per-topic.
Weighted versus unweighted

**GrMAW**

**GM – ignore weights**

- **Weighted average degree**
- **Average topic diameter**
- **#nodes**
Divide and conquer
Divide and conquer
Divide and conquer

- Divide problem based on $V$
Divide and conquer

- Divide problem based on $V$
- Conquer each sub-TCOs
- Divide problem based on \( V \)
- Conquer each sub-WTCoS
- Combine sub-WTCoS into one
Practicality of divide-and-conquer
Future work

- Overlay topologies + routing protocols
- Unstructured TCO + structured overlay
  - generalize TCO to be “interest-close”, nodes with similar interests reside close to each other
  - combine small-world network for navigability