BA: Constructing Fault-Tolerant Overlays for Topic-based Pub/Sub

Chen Chen
joint work with Roman Vitenberg and Hans-Arno Jacobsen
**Topic-connected overlay (TCO)** [PODC’07]

An overlay $G$

$G_a$ is topic-connected with one **TC-component**

$G_b$ is NOT topic-connected with two **TC-components**
MinAvg-$k$TCO problem

- $k$-topic-connected overlay ($k$TCO)
  Topic-connectivity still holds as long as fewer than $k$ nodes fail simultaneously on a topic.

- MinAvg-$k$TCO
  Given a set of nodes $V$, a set of topics $T$, and the interest function $Int$, construct a $k$TCO that has the least possible total number of edges, i.e., the minimum average node degree.
Our contributions

- Proposed MinAvg-$k$ TCO problem and analyzed its complexity
- Designed two algorithms
  - GM$2$ algorithm for MinAvg-$2$ TCO
  - HararyPT heuristic for MinAvg-$k$ TCO
- Conducted theoretical analysis and empirical evaluations
Complexity of MinAvg-\(k\)TCO

**Theorem:** For any positive integer \(k\), the MinAvg-\(k\)TCO problem parameterized by \(k\) is \textbf{NP-complete} and can not be approximated in polynomial time within a factor of \(O(\log |V|)\) unless \(P = NP\).
GM₂ algorithm for MinAvg-₂TCO

- **Lemma**: GM₂ outputs a ₂TCO within time
  \[ O(|V|^4|T|) \]

- **Lemma**: The approximation ratio is
  \[ O(U + \log |V||T|), \text{ where } U = \max_{t \in T} \{|V^{(t)}|\} \]

**Note**: MinAvg-₂TCO cannot be approximated within \( O(\log |V|) \) unless P = NP
Challenges to design GM²

Existing algorithms for TCO do not directly apply

- find progress measure towards 2TCO
- estimate progress as algorithm proceeds
- compare output to unknown optimum
Evaluation: GM\textsubscript{2}, GM, CyclePT

\[ \overline{d}_{\text{GM2}} \approx 1.5 \cdot \overline{d}_{\text{GM}} \]
Evaluation: HararyPT

![Graph showing the average node degree against the number of nodes for different models and values of k.](image)
BACKUP
GM₂ algorithm for MinAvg-$\mathcal{Z}$TCO

**Input:** $V, T, Int$

**Output:** $2TCO(V, T, Int, E_{GM₂})$

1. $E_{GM₂} \leftarrow \emptyset$
2. while $(V, T, Int, E_{GM₂})$ is not $2TCO$ do
3. $e \leftarrow$ find edge with maximum edge estimate
4. $E_{GM₂} \leftarrow E_{GM₂} \cup \{e\}$
5. return $2TCO(V, T, Int, E_{GM₂})$
HararyPT heuristic for MinAvg-$k$TCO

Input: (V, T, Int), k

Output: $k$TCO(V, T, Int, $E_{HPT}$)

1. V ← get an arbitrary sequence for V
2. for all $t \in T$ do
3. \[ E^{(t)} \leftarrow \text{buildHarary}(k, V^{(t)}) \] \[ \mathcal{O}(k|V||T|) \]
4. \[ E_{HPT} \leftarrow \bigcup_{t \in T} E^{(t)} \]
5. return $E_{HPT}$
Evaluation: Topic diameters

\[ \text{Diam}_{GM_2} \approx 0.5 \cdot \text{Diam}_{GM} \]
Two components in pub/sub implementation

- Construction of overlay
- Design of routing protocols

- Chockler et al. PODC'07
- Onus et al. INFOCOM'09
- G. Li et al. ICDCS'08
- M. Castro et al. JSAC'02

- Efficiently across the overlay network.
Why overlay for pub/sub?

- Limitations of other alternatives:
  - Full-mesh solution [GooPS; PNUTS]
    - Expensive communication overhead; Not scalable
  - Spanning tree
    - Vulnerable to churn; Costly to stabilize
Why overlay for pub/sub

- Advantages of the overlay approach:
  - Middleware philosophy and well-designed arch:
    - Decomposes the functionalities of routing protocols, overlay infrastructure, and network layer.
    - Reduces the complexity of the design and impl.
    - Enables exposing and virtualizing networking resources.
  - Demonstrates scalability and performance theoretically and empirically
    (mostly in topic-based pub/sub)
Why low node degrees?

- It costs lots of resources to maintain adjacent links for a high-degree node
  - A node must monitor the availability of each of its neighbors (heartbeats and keep-alive state)
- Each link takes responsibilities for a pub/sub
  - protocols
  - service components
  - message queues
- Other overlay designs also aim at low fan-out
  - DHT
  - Peer-to-peer video streaming
  - wireless network
Why topic-connected?

- **Eliminates immediate rely nodes:**
  
  Nodes not interested in a topic never need to contribute to message dissemination on that topic.

- **Supports efficient pub/sub routing**
  
  - Saves network and computational resources otherwise wasted on forwarding messages of no interest.
  
  - Results in simpler routing and smaller forwarding tables.

- **Provides better security**

  Messages are to be shared across a network among a set of trusted users without leaving this set.
Preliminaries

- $V$: a set of nodes
  - $T$: a set of topics
  - $Int$: an interest function: $V \times T \rightarrow \{\text{true, false}\}$
    - a node $v \in V$ is interested in some topic $t \in T$ iff $Int(v, t) = \text{true}$
  - $E$: set of output overlay edges

- Topic-connected component ($TC$-component): connected component in a topic-induced sub-graph for a topic $t$

- $TCO(V, T, Int, E)$: topic-connected overlay (TCO)
  - Contains only one $TC$-component for each topic