Consensus Community Detection for Multi-dimensional Networks

Antonela Tommasel

Daniela Godoy

ISISTAN
Instituto Superior de Ingenieria de Software Tandil

Table of Contents

Social Networks

- Social networks can be defined as a set of socially relevant nodes connected by one or more relations.

- Social media users have greater freedom to connect with a wider number of people for distinct reasons.

Social Networks

- Social networks can be defined as a set of socially relevant nodes connected by one or more relations.
- Social media users have greater freedom to connect with a wider number of people for distinct reasons.
- The pervasive use of social media offers research opportunities for analysing the behaviour of users when interacting with their friends.

Social Networks

- Social networks can be defined as a set of socially relevant nodes connected by one or more relations.
- Social media users have greater freedom to connect with a wider number of people for distinct reasons.
- The pervasive use of social media offers research opportunities for analysing the behaviour of users when interacting with their friends.

One fundamental problem in social networks is to identify groups of users, even when group information is not explicitly available!!

Community Detection

- A group, or community, can be defined as a set of elements that interact more frequently or share more similarity with other community members than with outsiders.

Community Detection

- A group, or community, can be defined as a set of elements that interact more frequently or share more similarity with other community members than with outsiders.

- Algorithms only focus on **one source of information**, even though **neither social relations nor content** alone can accurately indicate community membership.

Community Detection

- A group, or community, can be defined as a set of elements that interact more frequently or share more similarity with other community members than with outsiders.

- Algorithms only focus on **one source of information**, even though **neither social relations nor content** alone can accurately indicate community membership.

- Community detection techniques should combine **multiples sources of information**.

Community Detection

- **New Challenges!!**
  - **How to extract information belonging to multiple and heterogeneous information sources?**
  - **How to integrate the different information sources?**
  - **How to represent the graph?**
  - **How to perform community detection over heterogeneous graphs?**

Community Detection

- **New Challenges!!**
  - **How to extract information belonging to multiple and heterogeneous information sources?**
  - **How to integrate the different information sources?**
  - **How to represent the graph?**
  - **How to perform community detection over heterogeneous graphs?**

Community Detection

- **We propose…**
  - **Consider a multi-dimensional graph representation.**
  - **Present and analyse four integration techniques for applying traditional community detection techniques to multi-dimensional graphs.**

Community Detection

- **We propose…**
  - **Consider a multi-dimensional graph representation.**
  - **Present and analyse four integration techniques for applying traditional community detection techniques to multi-dimensional graphs.**

The final goal is to provide some insights on **how to integrate the diverse information sources** and **user interactions** for **improving** the **quality** of hidden **community structures** that are shared by the **heterogeneous interactions**.

Table of Contents

Consensus Community Detection

Graph Representation

Consensus Strategies

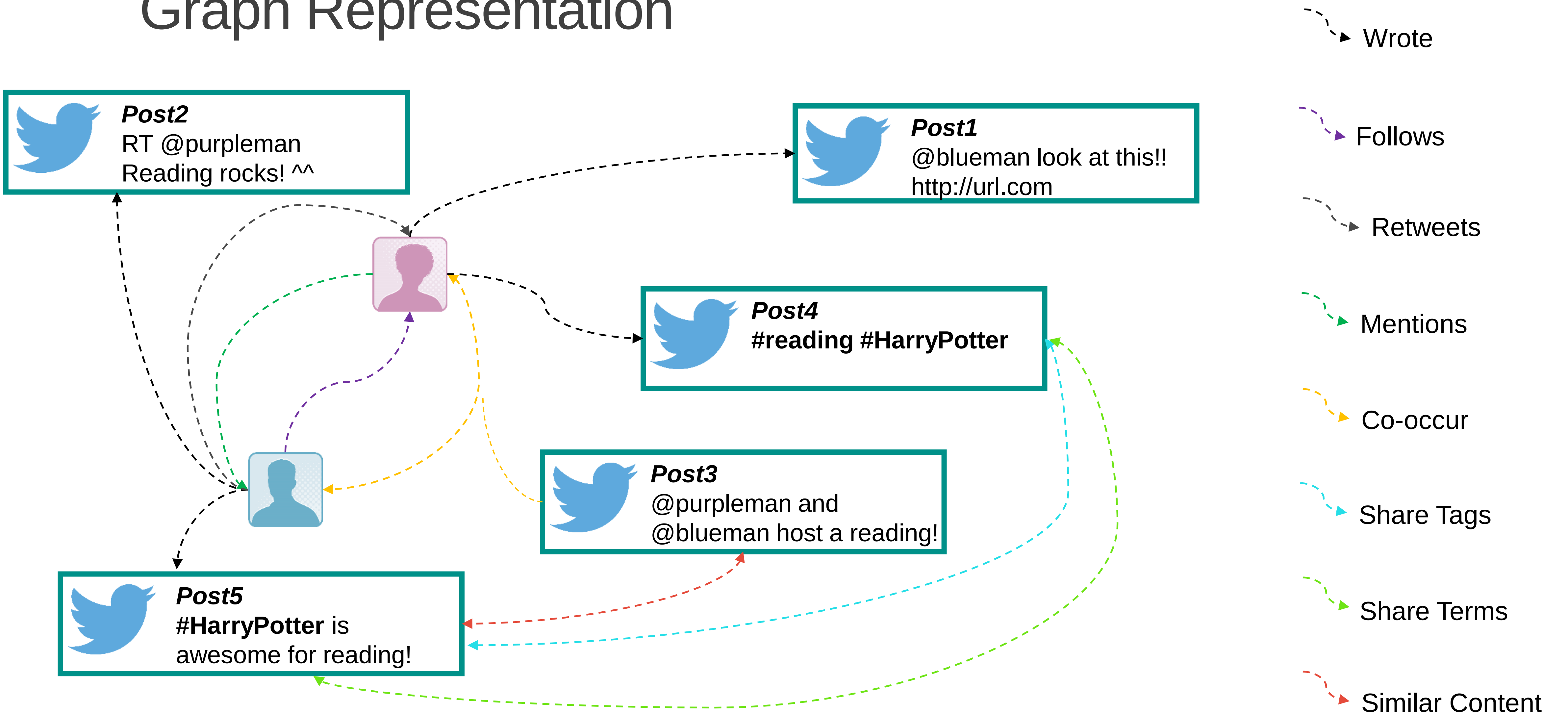
Consensus Community Detection

Graph Representation

- Information and content-based relations offer **complementary** views of data.
- **No** individual relation **alone** might be **sufficient** for accurately determining community membership.
- A social relation between two nodes was established if authors in a node followed authors of the other node.
- Different content-based relations were extracted from data.

Consensus Community Detection

Graph Representation



Consensus Community Detection

Graph Representation

Shared Tag	Shared Class	Similar Content

Consensus Community Detection

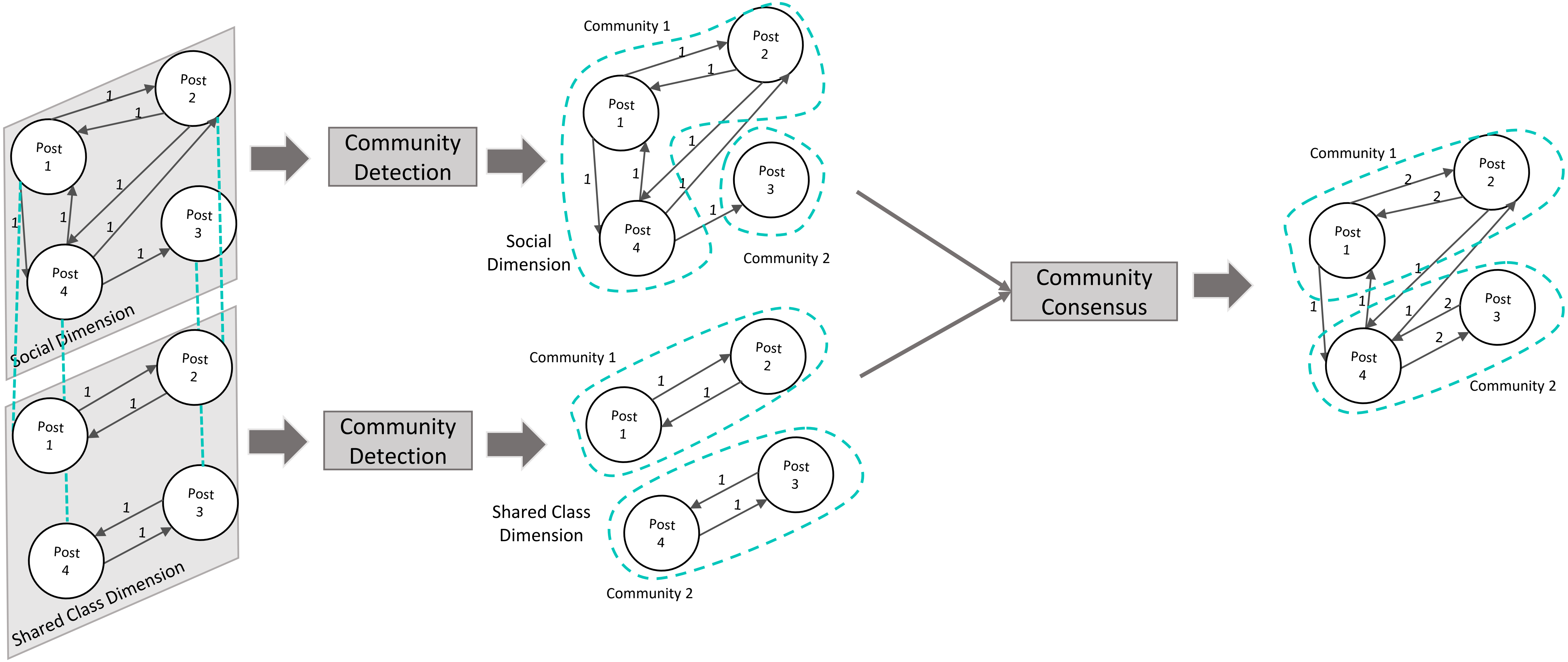
Graph Representation

Shared Tag	Shared Class	Similar Content

- *Shared Tags.* An edge between two nodes exists if they **share any tag (or hashtag)**. Edge weight is measured as the percentage of shared tags amongst the total number of different tags comprised by the two posts.
- *Shared Class.* An edge between two nodes exists if they **belong to the same class**. All edges have a weight of 1.
- *Similar Content.* Measures the **content resemblance of two nodes**. Edge weight is defined by means of the Cosine Similarity.

Consensus Community Detection

Consensus Strategies



Consensus Community Detection

Consensus Strategies

Instance based	Cluster based
Hybrid bipartite	Metric based

Consensus Community Detection

Consensus Strategies

Instance based

- Each data instance is represented as a node in the new graph.
- Edges are weighted proportionally to how frequently the two nodes are located in the same community.

- Three weighting alternatives:
  - *Instance-based:* the number of shared communities.
  - *Instance-based-%:* the percentage of shared communities.
  - *Instance-based-content:* amongst the total number of communities the number of shared communities scaled by the nodes’ content similarity.

- The partition resulting from the new graph is the final community partition.

Consensus Community Detection

Consensus Strategies

- Each community in each partition is represented as a node in the new graph.
- Edges are weighted according to the Jaccard Similarity considering the data instances in each community.

- The graph is partitioned and each group of communities represents a meta-community.

- Each original data instance is assumed to be associated to a meta-community if such meta-community includes a community the data instance belongs to.

- Assumes that there exists a structural correspondence amongst the different community partitions.

Cluster based

Consensus Community Detection

Consensus Strategies

Hybrid bipartite

- Creates a bipartite graph, in which nodes represent both the original data instances and the communities.
- Edges only connect nodes representing data instances with nodes representing communities, according to whether the data instance is included in the community.
- All edges have weight 1.

Consensus Community Detection

Consensus Strategies

- Does not combine the different partitions.
- Assesses the quality of the different partitions in terms of a quality metric.

- The partition achieving the highest quality results is selected as the final partition.

- Quality is assessed by two metrics.
  - *Metric-based-betweenness:* betweenness centrality.
  - *Metric-based-content:* the average content similarity of communities.

Metric based

Table of Contents

Dataset

- Experimental evaluation was based on Twitter.
- Dataset included more than 500,000 tweets classified into 1,036 trending topics.
- Each trending topic was considered as a node in the graph.

Number of Instances	1,036
Number of Features	226,043
Number of Classes	4
Number of Following Relations	251,522,840
Average number of Followees	816
Average number of Features per Instance	1,084
Average number of Instances per Class	259

Experimental Settings

- Strategies were evaluated considering the Gephi implementation of the Louvain algorithm.
- Two experimental settings were considered:
  - Content and social relations are independent (each relation can create new edges).
  - Content features are used to weight the social relations.
- Community quality was evaluated by two types of scoring functions.
  - Functions characterising the connectivity structure of communities.
  - Functions characterising communities’ content cohesiveness.

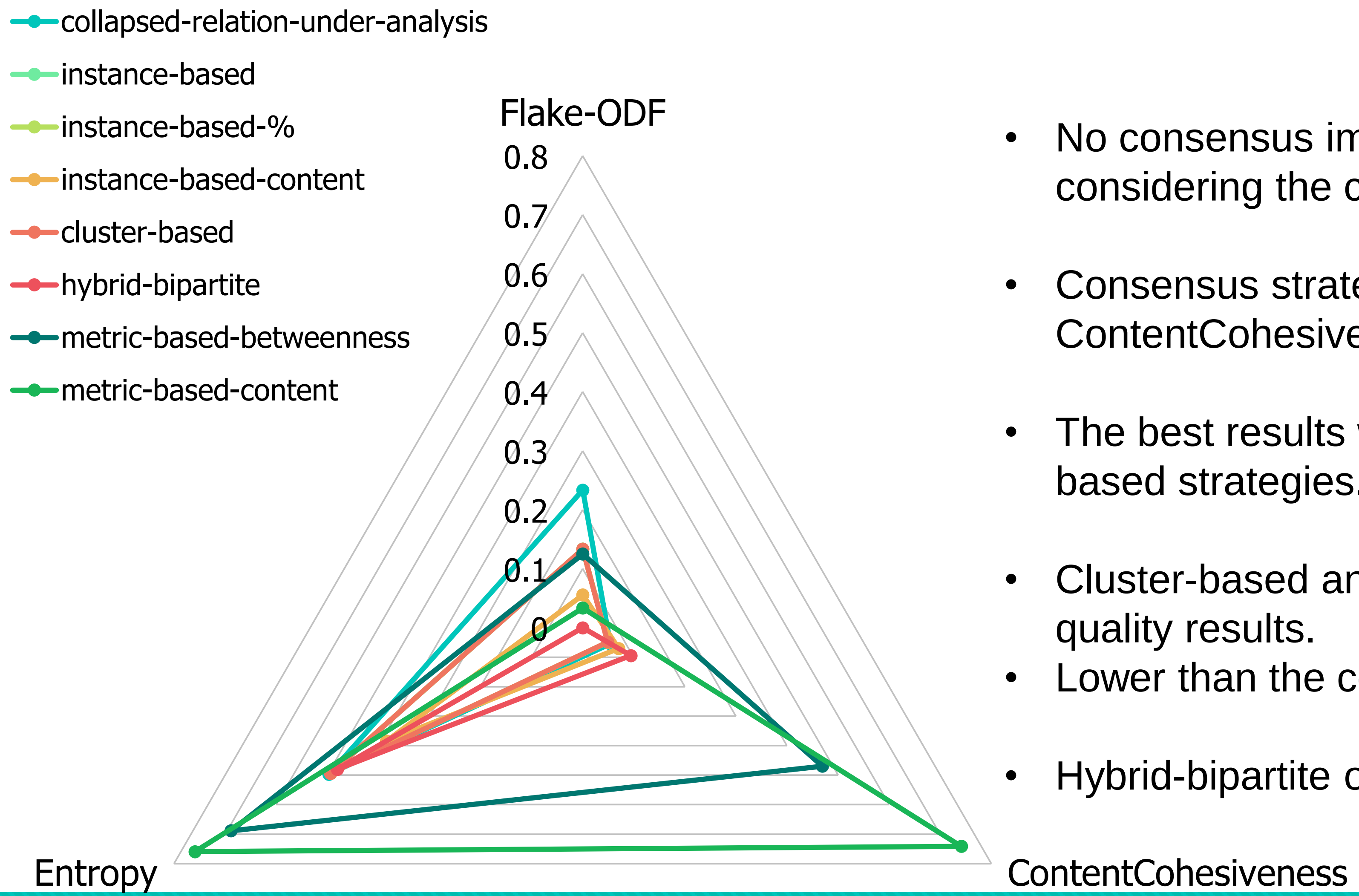
Experimental Settings

Treating Social and Content relations independently
Social & SimilarContent-0.6
Social & SharedClass
Social & SharedClass & SharedTag & SimilarContent-0.6
Social & SharedClass & SharedTag & SimilarContent

Weighting Social relations with content-based relations
Social-W-SimilarContent-0.6 & SharedClass
Social-W-SharedClass & SimilarContent-0.6
Social-W-SharedTag & SharedClass

Experimental Results

Independent Social and Content Views

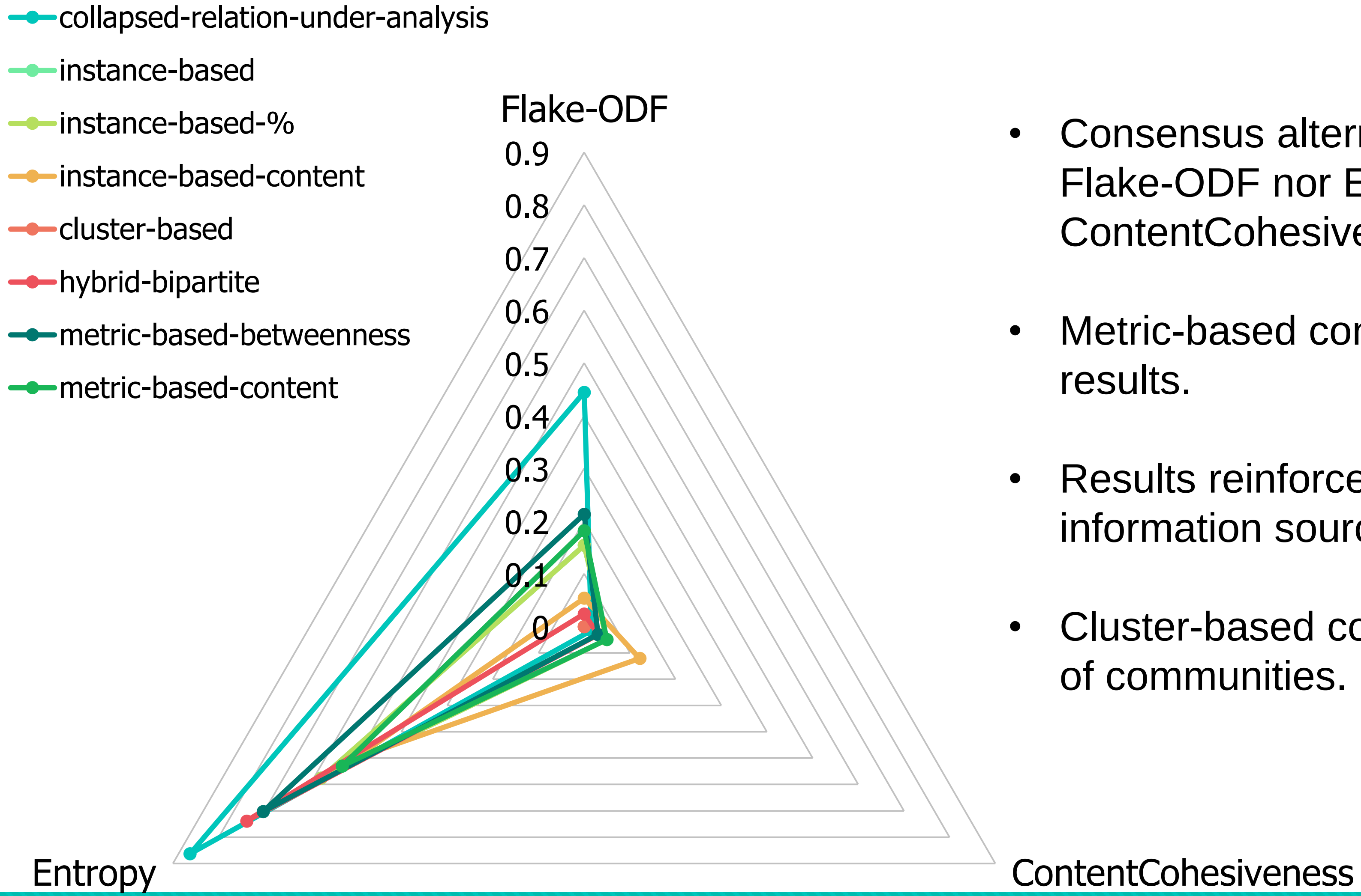


Social & Similar-Content-0.6

- No consensus improved the Flake-ODF results of considering the collapsed graph representation
- Consensus strategies improved Entropy and ContentCohesiveness.
- The best results were obtained when considering the metric based strategies.
- Cluster-based and instance-based strategies obtained similar quality results.
- Lower than the collapsed graph.
- Hybrid-bipartite obtained the worst results.

Experimental Results

Independent Social and Content Views

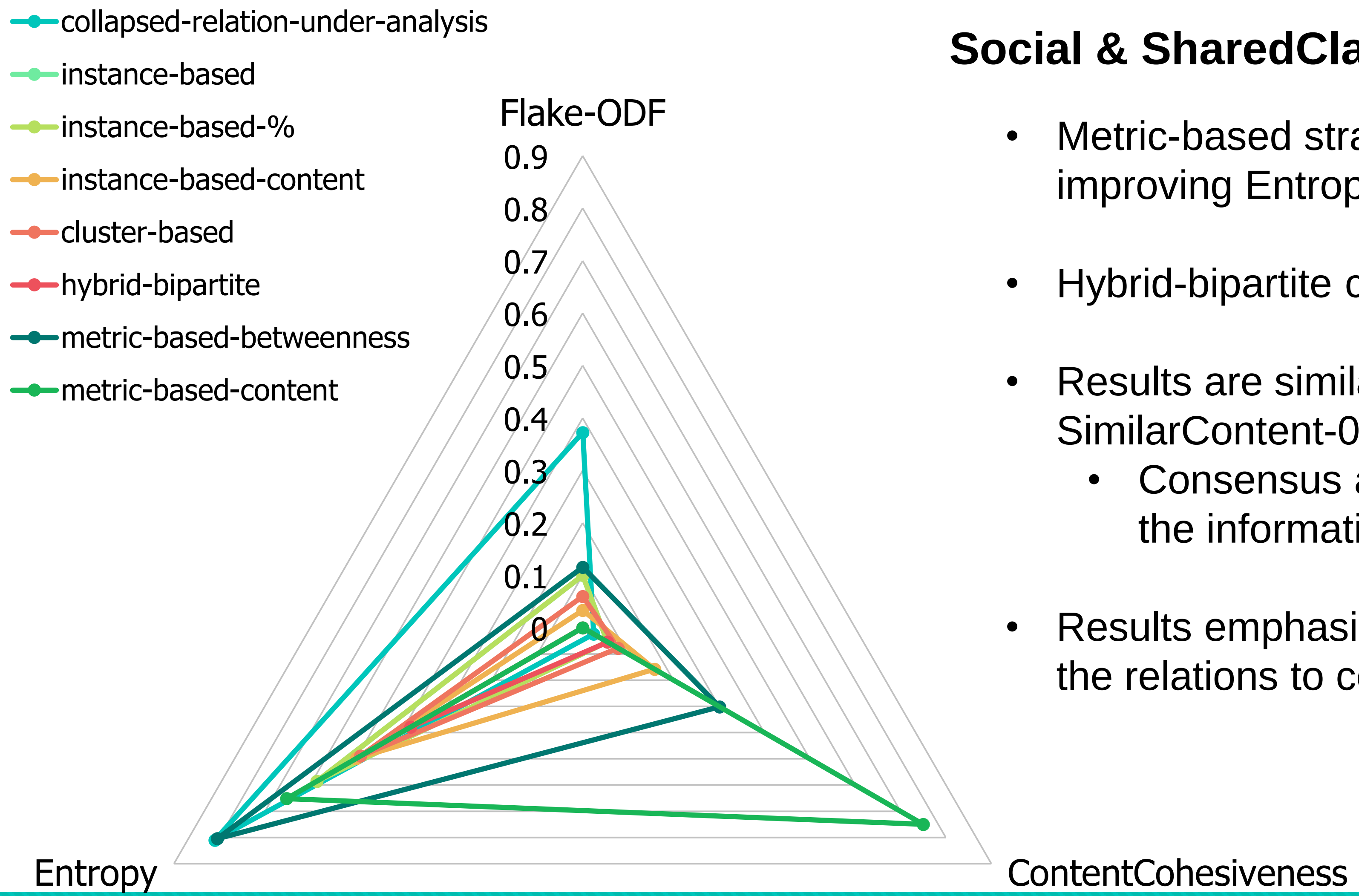


Social & SharedClass

- Consensus alternatives were not able to improve neither Flake-ODF nor Entropy of the collapsed graph, but improved ContentCohesiveness.
- Metric-based consensus alternatives obtained the best results.
- Results reinforced the complementary nature of the diverse information sources.
- Cluster-based consensus did not find a meaningful number of communities.

Experimental Results

Independent Social and Content Views

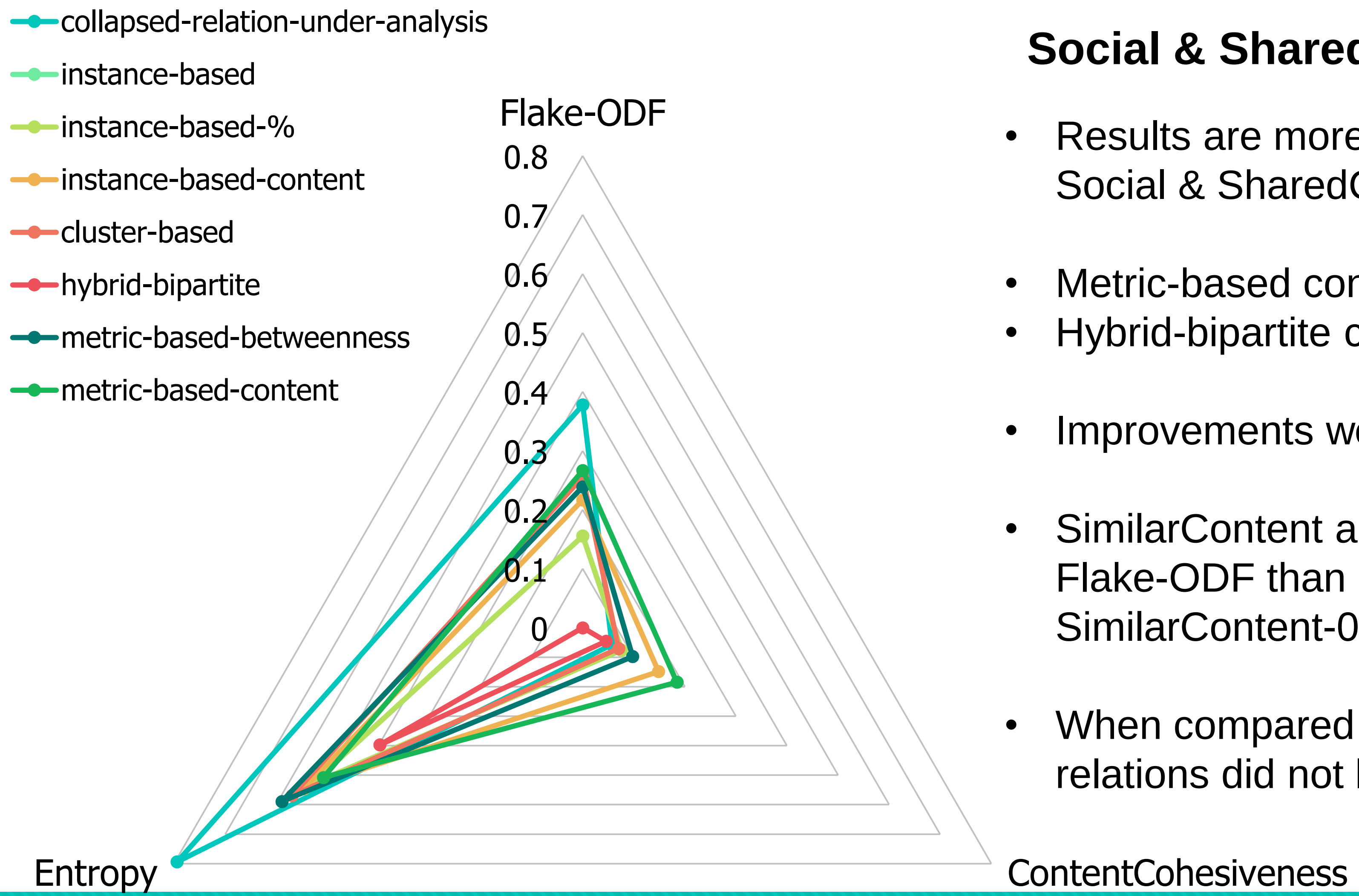


Social & SharedClass & SharedTag & SimilarContent-0.6

- Metric-based strategies found highest quality communities, improving Entropy and ContentCohesiveness.
- Hybrid-bipartite consensus obtained the worst results.
- Results are similar as those obtained for Social & SimilarContent-0.6.
  - Consensus alternatives were unable to leverage on all the information provided by every analysed relation.
- Results emphasise the importance of adequately choosing the relations to consider.

Experimental Results

Independent Social and Content Views

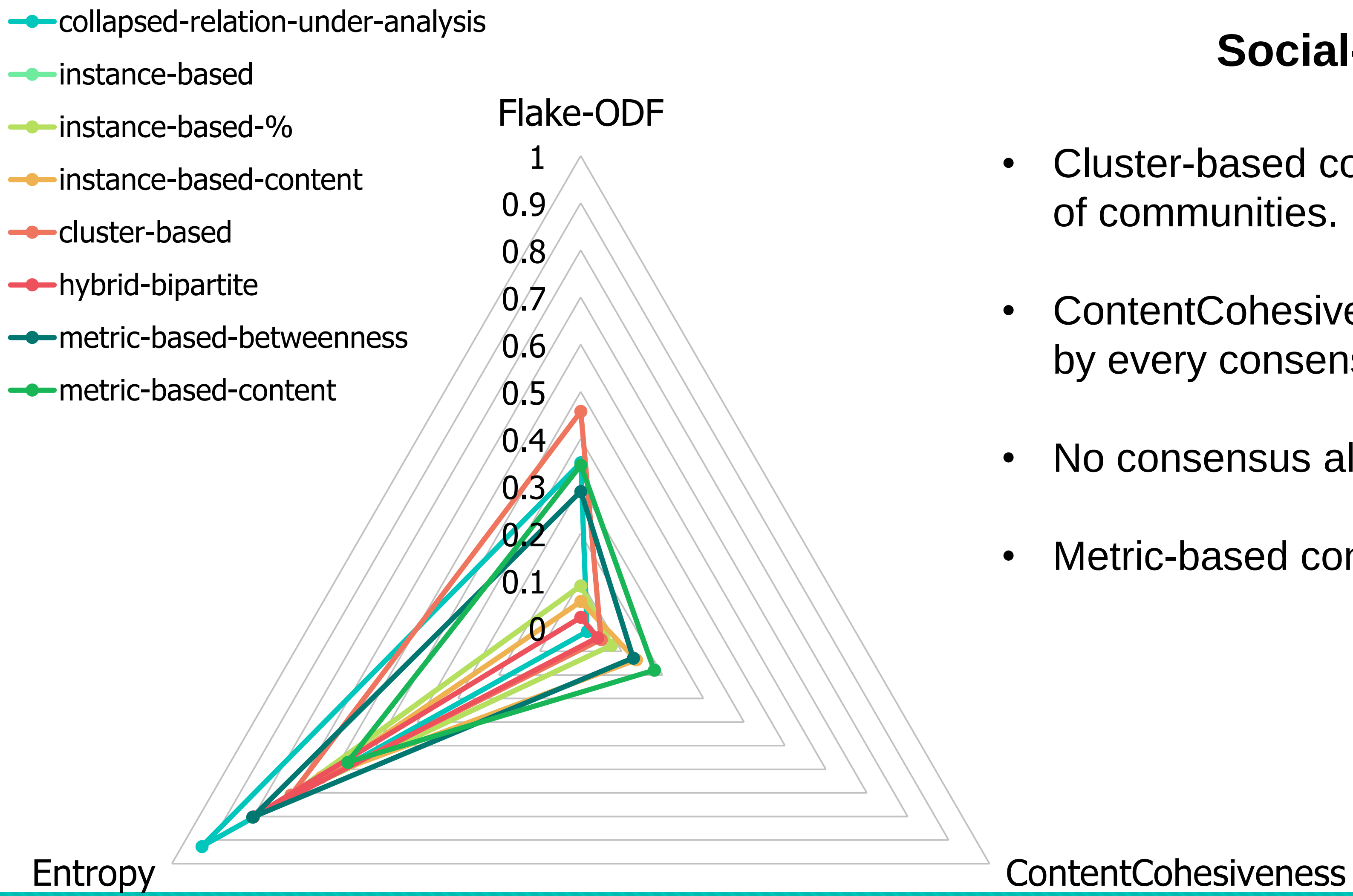


Social & SharedClass & SharedTag & SimilarContent

- Results are more similar to Social & SharedClass than to Social & SharedClass & SharedTag & SimilarContent-0.6.
- Metric-based consensus obtained the best results.
- Hybrid-bipartite consensus obtained the worst results.
- Improvements were lower than for SimilarContent-0.6.
- SimilarContent allowed to find communities with higher Flake-ODF than Social & SharedClass & SharedTag & SimilarContent-0.6.
- When compared to Social & SharedClass, adding more relations did not lead to better Entropy results.

Experimental Results

Weighting Social View

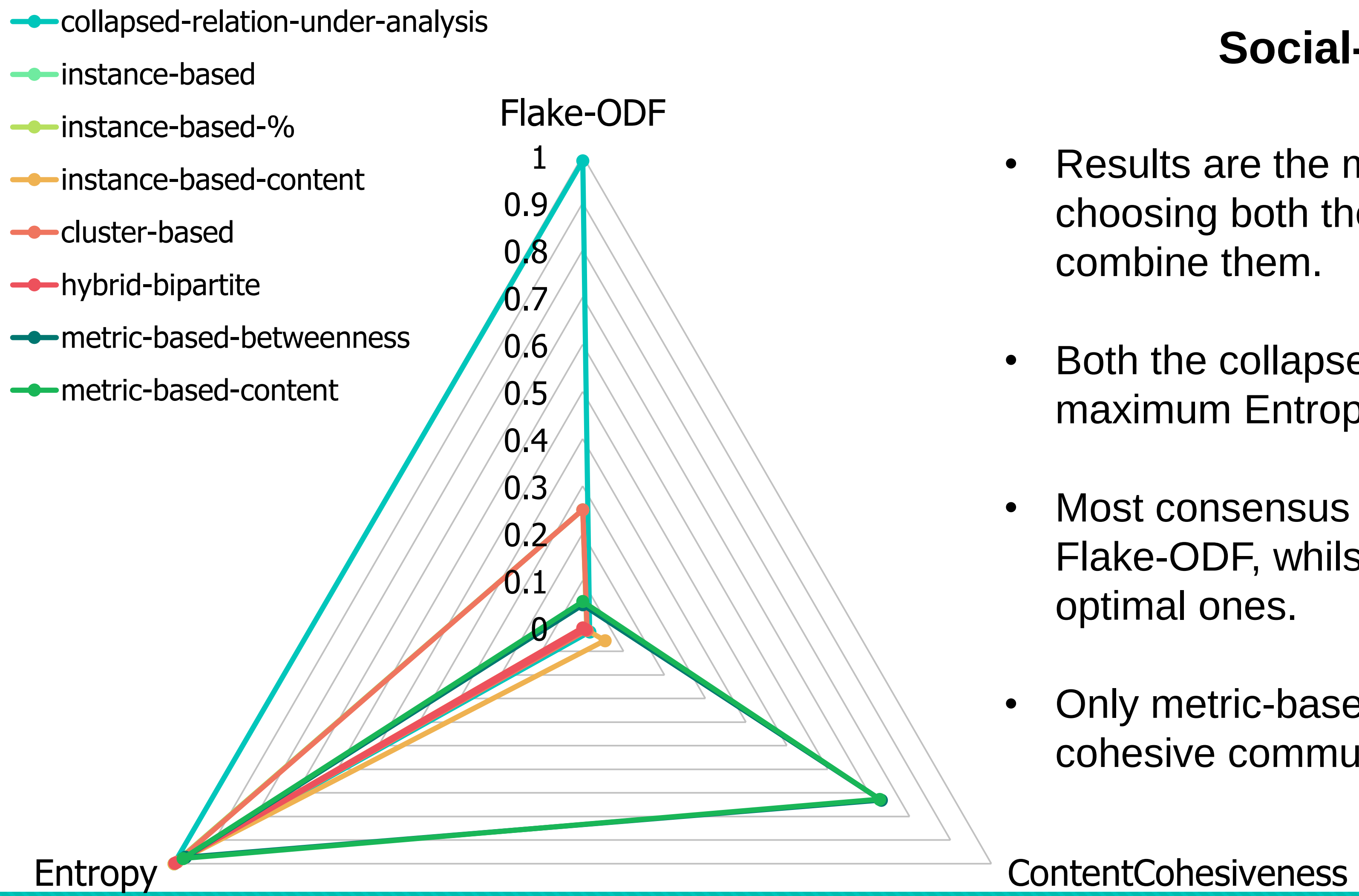


Social-W-SimilarContent-0.6 & SharedClass

- Cluster-based consensus strategy improved the Flake-ODF of communities.
- ContentCohesiveness of the collapsed graph were improved by every consensus strategy.
- No consensus alternatives was able to improve Entropy.
- Metric-based consensus obtained the best overall results.

Experimental Results

Weighting Social View

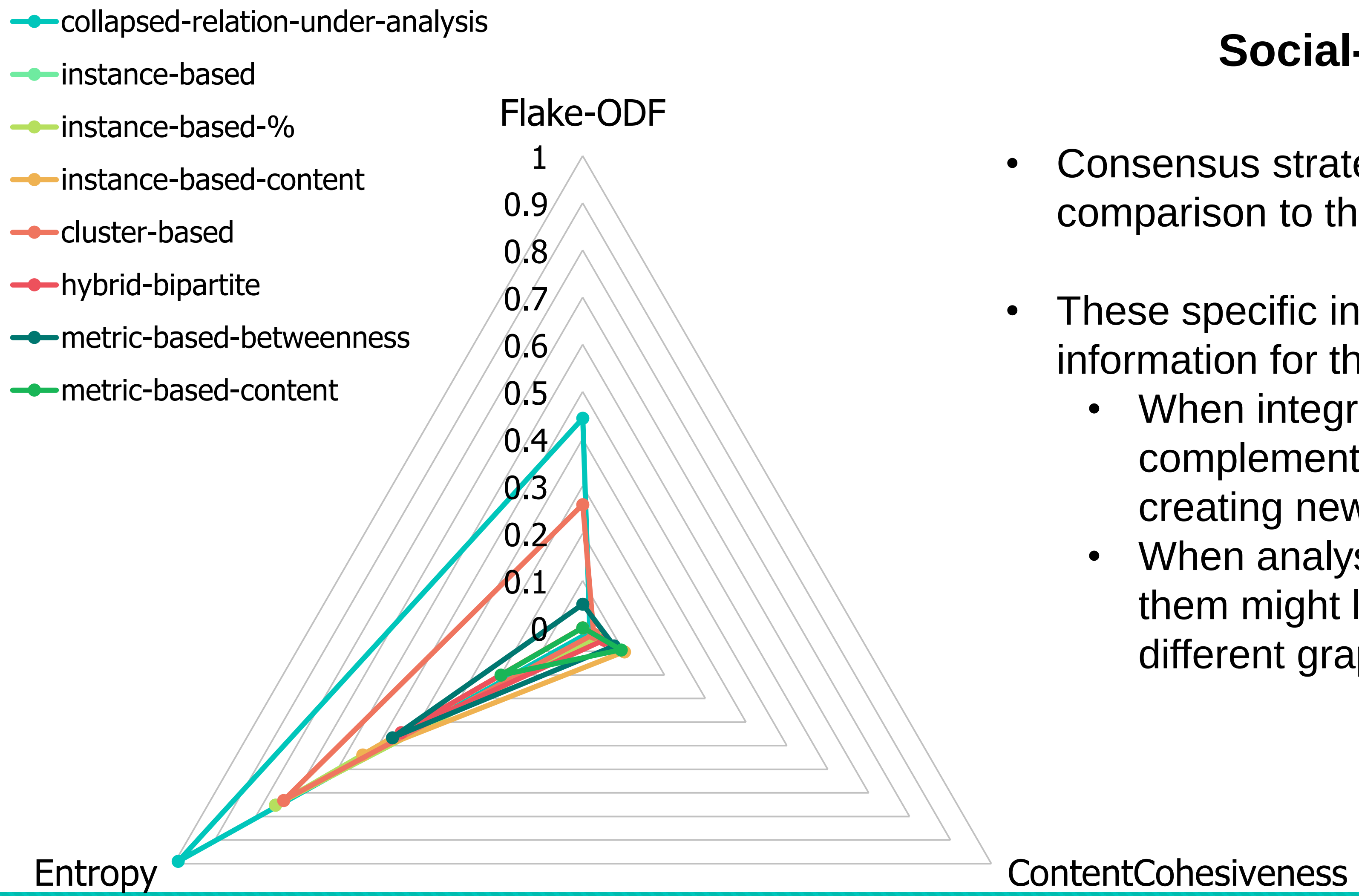


Social-W-SharedClass & SimilarContent-0.6

- Results are the most representative of the importance of choosing both the adequate node relations, and how to combine them.
- Both the collapsed graph and consensus strategies achieved maximum Entropy.
- Most consensus strategies obtained close to zero results for Flake-ODF, whilst the collapsed representation obtained the optimal ones.
- Only metric-based consensus was able to find high content cohesive communities.

Experimental Results

Weighting Social View



Social-W-SimilarContent-0.6 & SharedClass

- Consensus strategies obtained the worst results in comparison to the other combinations of relationships.
- These specific individual relations do not provide enough information for the community detection algorithm.
  - When integrated into the same graph, a cohesive and complementary graph is created by strengthening or creating new links between nodes.
  - When analysing each relation individually, each of them might lead to sparse graphs or even completely different graph structures.

Experimental Results

Summary of Results

- **Metric-based consensus strategies obtained in all cases the highest quality partitions.**

- **Hybrid-bipartite consensus obtained the worst community partitions.**

Experimental Results

Summary of Results

- **Metric-based consensus strategies obtained in all cases the highest quality partitions.**
- **Hybrid-bipartite consensus obtained the worst community partitions.**
- Using consensus strategies to combine the communities found by individual node relations **did not always** yield the highest quality results.

Experimental Results

Summary of Results

- **Metric-based consensus strategies obtained in all cases the highest quality partitions.**
- **Hybrid-bipartite consensus obtained the worst community partitions.**
- Using consensus strategies to combine the communities found by individual node relations **did not always** yield the highest quality results.
- Results showed that for **some** combinations of relations, the highest quality communities were obtained when **collapsing** relations into a unique graph.

Experimental Results

Summary of Results

- **Metric-based consensus strategies obtained in all cases the highest quality partitions.**

- **Hybrid-bipartite consensus obtained the worst community partitions.**

- Using consensus strategies to combine the communities found by individual node relations **did not always** yield the highest quality results.

- Results showed that for **some** combinations of relations, the highest quality communities were obtained when **collapsing** relations into a unique graph.

Why?

Experimental Results

Summary of Results

- The information provided by each individual relation **might not be neither enough nor accurate** for discovering the community structure.
- Relations might be **redundant**. Adding more relations does not necessarily imply adding new information to the process.
- Relations might provide **contradictory** information regarding the nodes.

Table of Contents

Summary

- This work aimed at **analysing several consensus strategies** for extending community detection techniques designed for a unique data dimension to multi-dimensional networks.

Summary

- This work aimed at **analysing several consensus strategies** for extending community detection techniques designed for a unique data dimension to multi-dimensional networks.

- Tackled the problem of how to **combine diverse information sources** available in social media data to optimise the quality of the discovered communities.

Summary

- This work aimed at **analysing several consensus strategies** for extending community detection techniques designed for a unique data dimension to multi-dimensional networks.

- Tackled the problem of how to **combine diverse information sources** available in social media data to optimise the quality of the discovered communities.

- **Showed the effect that each consensus strategy has over community quality.**

Summary

What to choose? Collapsed or Consensus?

- The decision should be **guided by the characteristics of the data under analysis**.
  - As Twitter is a social platform aimed at sharing information, content-based relations might convey more information than the social ones.

Summary

What to choose? Collapsed or Consensus?

- The decision should be **guided by the characteristics of the data under analysis**.
    - As Twitter is a social platform aimed at sharing information, content-based relations might convey more information than the social ones.

- When **more than two** relations are meant to be used, it is **likely** that the **same results** would be obtained with **fewer relations** given the selection of the adequate consensus strategy.

Summary

What to choose? Collapsed or Consensus?

- The decision should be **guided by the characteristics of the data under analysis**.
  - As Twitter is a social platform aimed at sharing information, content-based relations might convey more information than the social ones.

- When **more than two** relations are meant to be used, it is **likely** that the **same results** would be obtained with **fewer relations** given the selection of the adequate consensus strategy.

- When relations might be **contradictory**, it might be preferable to **collapse them into a unique** graph, as the nature of relations might mislead the consensus strategy.

Conclusions

- **Consensus strategies** could help to **improve the quality of communities** with respect to collapsing multiple heterogeneous relations into a unique graph.

- The diverse **consensus strategies** have a **distinct effect** over the quality of communities.

Conclusions

- **Consensus strategies** could help to **improve the quality of communities** with respect to collapsing multiple heterogeneous relations into a unique graph.
- The diverse **consensus strategies** have a **distinct effect** over the quality of communities.
- Diverse **information sources** were found **not to evenly contribute** to the improvement of community quality.
- The **behaviour** of the information sources **differs** according to whether they are **mixed together or individually analysed.**

Questions?

Consensus Community Detection for Multi-dimensional Networks

Antonela Tommasel

Daniela Godoy

ISISTAN
Instituto Superior de Ingenieria de Software Tandil