Hyper-Flow Diffusion

Kimon Fountoulakis\textsuperscript{1}, Pan Li\textsuperscript{2}, Shenghao Yang\textsuperscript{1}

\textsuperscript{1}University of Waterloo \quad \textsuperscript{2}Purdue University

Networks 2021
Hypergraphs generalize graphs by allowing a hyperedge to consist of multiple nodes that capture higher-order relations in the data.

**E-commerce**
Nodes are products or webpages
Several products can be purchased at once
Several webpages are visited during the same session

**Collaboration**
Nodes are authors
A group of authors collaborate on a paper/project

**Ecology**
Nodes are species
Multiple species interact according to their roles in the food chain
Diffusion algorithms are everywhere (for graphs)

About 4,240,000 results (0.04 sec)
Diffusion algorithms are everywhere (for graphs)

Diffusion on a graph is the process of spreading a given initial mass from some seed node(s) to neighbor nodes using the edges of the graph.

Applications include recommendation systems, node ranking, community detection, social and biological network analysis, etc.
Diffusion algorithms are everywhere (for graphs)

However ... hypergraph diffusion has been significantly less explored:
Existing methods either do not have a tight theoretical implication, or do not model complex high-order relations, or are not scalable to large datasets.
Our motivation

We propose the first local diffusion method that

• Achieves **stronger theoretical guarantees** for the local hypergraph clustering problem;

• Applies to a **substantially richer class of higher-order relations** with only a submodularity assumption;

• Permits **computational efficient** algorithms.

**However … hypergraph diffusion has been significantly less explored:** Existing methods either do not have a **tight theoretical implication**, or do not model **complex high-order relations**, or are not **scalable** to large datasets.
Higher-order relations: hyperedge cut perspective

There are distinct ways to cut a 4-node hyperedge.

How do we treat \(v_1\) \(v_2\) \(v_3\) \(v_4\) differently from \(v_1\) \(v_2\) \(v_3\) \(v_4\)?
Higher-order relations: hyperedge cut perspective

Distinct ways to cut a 4-node hyperedge may have different costs.

$w_e(S)$ specifies the cost of splitting $e$ into $S$ and $e \setminus S$. 
Higher-order relations: hyperedge cut perspective

Distinct ways to cut a 4-node hyperedge may have different costs.

Unit: the cost of cutting a hyperedge is always 1, i.e., $w_e(S) = 1$

$w_e(\{v_2\}) = 1$

$w_e(\{v_1, v_2\}) = 1$

$w_e(\{v_1, v_3\}) = 1$

$w_e(S)$ specifies the cost of splitting $e$ into $S$ and $e \setminus S$. 
Higher-order relations: hyperedge cut perspective

Distinct ways to cut a 4-node hyperedge may have different costs.

**Unit:** the cost of cutting a hyperedge is always 1, i.e., \( w_e(S) = 1 \).

**Cardinality-based:** the cost of cutting a hyperedge depends on the number of nodes in either side of the hyperedge, i.e., \( w_e(S) = f(\min\{ |S|, |e \setminus S| \}) \).

\( w_e(S) \) specifies the cost of splitting \( e \) into \( S \) and \( e \setminus S \).
Higher-order relations: hyperedge cut perspective

Distinct ways to cut a 4-node hyperedge may have different costs.

Unit: the cost of cutting a hyperedge is always 1, i.e., $w_e(S) = 1$.

Cardinality-based: the cost of cutting a hyperedge depends on the number of nodes in either side of the hyperedge, i.e., $w_e(S) = f(\min\{ |S|, |E \setminus S| \})$.

Submodular: the costs of cutting a hyperedge form a submodular function, i.e., $w_e : 2^e \to \mathbb{R}$ is a submodular set function.

$w_e(S)$ specifies the cost of splitting $e$ into $S$ and $E \setminus S$. 

$w_e(\{v_2\}) = 1$

$w_e(\{v_1, v_2\}) = 0$

$w_e(\{v_1, v_3\}) = 2$
A food network can be mapped into a hypergraph by taking each network pattern on the left as a hyperedge on the right. This network pattern captures carbon flow from two preys ($v_1, v_2$) to two predators ($v_3, v_4$).
Higher-order relations: hyperedge cut perspective

The cut-cost $w_e(\{v_1, v_2\}) = w_e(\{v_3, v_4\}) = 0$ encourages separation of predators and preys.
Higher-order relations: hyperedge cut perspective

The cut-cost $w_e(\{v_1, v_2\}) = w_e(\{v_3, v_4\}) = 0$ encourages separation of predators and preys.

The cut-cost $w_e(\{v_1, v_3\}) = w_e(\{v_2, v_4\}) = 2$ discourages grouping of predators and preys.
Higher-order relations: hyperedge cut perspective

The cut-cost \( w_e(\{v_1, v_2\}) = w_e(\{v_3, v_4\}) = 0 \) encourages separation of predators and preys.

The cut-cost \( w_e(\{v_1, v_3\}) = w_e(\{v_2, v_4\}) = 2 \) discourages grouping of predators and preys.

The cut-cost \( w_e(\{v_1\}) = w_e(\{v_2\}) = w_e(\{v_3\}) = w_e(\{v_4\}) = 1 \) assigns less penalty for separating a single node. It also makes \( w_e : 2^e \rightarrow \mathbb{R}_+ \) a submodular function.
Higher-order relations: hyperedge flow perspective

To specify flows (i.e., movement of mass) over an edge or hyperedge, we associate each node a number which indicates the direction (sign) and magnitude of flow.
Higher-order relations: hyperedge flow perspective

Flows on graph

Flows on hypergraph

A natural generalization of network flows.

Flow conservation: numbers within the same hyperedge sum to 0.
Additional constraints required for hyperedges so that the numbers reflect higher-order relations.
Hyper-Flow Diffusion

- Initial mass $\Delta$ on some seed node(s)

$\Delta = 5$
Hyper-Flow Diffusion

- Initial mass $\Delta$ on some seed node(s)
- Diffuse mass according to flows over hyperedges
Hyper-Flow Diffusion

- Initial mass $\Delta$ on some seed node(s)
- Diffuse mass according to flows over hyperedges
- Leave net mass $m$ on nodes
Hyper-Flow Diffusion

- Initial mass $\Delta$ on some seed node(s)
- Diffuse mass according to flows over hyperedges
- Leave net mass $m$ on nodes
- Net mass cannot exceed capacity $d$
Hyper-Flow Diffusion

- Initial mass $\Delta$ on some seed node(s)
- Diffuse mass according to flows over hyperedges
- Leave net mass $m$ on nodes
- Net mass cannot exceed capacity $d$

We impose additional constraints so that the flow values respect higher-order relations modelled by the cut-cost function $w_e$.

Hyper-Flow Diffusion is the diffusion of initial mass according to minimum $\ell_2$-norm flow.
Hyper-Flow Diffusion

- Initial mass $\Delta$ on some seed node(s)
- Diffuse mass according to flows over hyperedges
- Leave net mass $m$ on nodes
- Net mass cannot exceed capacity $d$

We impose additional constraints so that the flow values respect higher-order relations modelled by the cut-cost function $w_e$.

Hyper-Flow Diffusion is the diffusion of initial mass according to minimum $\ell_2$-norm flow.

We use the excess mass on nodes for node ranking and local clustering
Hyper-Flow Diffusion: empirical results

Cardinality-based $k$-uniform hypergraph stochastic block model:
Boundary hyperedges appear with different probabilities according to
the cardinality of hyperedge cut.

We consider $q_1 \gg q_2 \geq q_3$. Under this generative setting, one should
naturally explore cardinality-based cut-cost for clustering.

All our experiments use a single seed node to recover the target.
Hyper-Flow Diffusion: empirical results

- LH is a strongly-local hypergraph diffusion method based on graph reduction.
- ACL is a heuristic method that uses PageRank on star expansion.
- HFD is the only method that directly works on original hypergraph.
- U-* means the method uses unit cut-cost; C-* means the method uses cardinality cut-cost.
- For each method, C-* is better than U-*.
- There is a significant performance drop for C-LH at $k = 4$. 
Hyper-Flow Diffusion: empirical results

Local clustering on a hypergraph constructed from Amazon product reviews data

<table>
<thead>
<tr>
<th>Metric</th>
<th>Seed</th>
<th>Method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>12</th>
<th>15</th>
<th>17</th>
<th>18</th>
<th>24</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conductance Single</td>
<td></td>
<td><strong>U-HFD</strong></td>
<td><strong>0.17</strong></td>
<td><strong>0.11</strong></td>
<td><strong>0.12</strong></td>
<td><strong>0.16</strong></td>
<td><strong>0.36</strong></td>
<td><strong>0.25</strong></td>
<td><strong>0.17</strong></td>
<td><strong>0.14</strong></td>
<td><strong>0.28</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>U-LH-2.0</strong></td>
<td><strong>0.42</strong></td>
<td><strong>0.50</strong></td>
<td><strong>0.25</strong></td>
<td><strong>0.44</strong></td>
<td><strong>0.74</strong></td>
<td><strong>0.44</strong></td>
<td><strong>0.57</strong></td>
<td><strong>0.58</strong></td>
<td><strong>0.61</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>U-LH-1.4</strong></td>
<td><strong>0.33</strong></td>
<td><strong>0.44</strong></td>
<td><strong>0.25</strong></td>
<td><strong>0.36</strong></td>
<td><strong>0.81</strong></td>
<td><strong>0.40</strong></td>
<td><strong>0.51</strong></td>
<td><strong>0.54</strong></td>
<td><strong>0.59</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>ACL</strong></td>
<td><strong>0.42</strong></td>
<td><strong>0.50</strong></td>
<td><strong>0.25</strong></td>
<td><strong>0.54</strong></td>
<td><strong>0.77</strong></td>
<td><strong>0.52</strong></td>
<td><strong>0.63</strong></td>
<td><strong>0.68</strong></td>
<td><strong>0.65</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>U-HFD</strong></td>
<td><strong>0.05</strong></td>
<td><strong>0.10</strong></td>
<td><strong>0.12</strong></td>
<td><strong>0.13</strong></td>
<td><strong>0.20</strong></td>
<td><strong>0.16</strong></td>
<td><strong>0.14</strong></td>
<td><strong>0.11</strong></td>
<td><strong>0.32</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>U-LH-2.0</strong></td>
<td><strong>0.05</strong></td>
<td><strong>0.15</strong></td>
<td><strong>0.15</strong></td>
<td><strong>0.21</strong></td>
<td><strong>0.45</strong></td>
<td><strong>0.45</strong></td>
<td><strong>0.26</strong></td>
<td><strong>0.18</strong></td>
<td><strong>0.53</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>U-LH-1.4</strong></td>
<td><strong>0.05</strong></td>
<td><strong>0.13</strong></td>
<td><strong>0.15</strong></td>
<td><strong>0.15</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.33</strong></td>
<td><strong>0.19</strong></td>
<td><strong>0.14</strong></td>
<td><strong>0.47</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>ACL</strong></td>
<td><strong>0.05</strong></td>
<td><strong>0.27</strong></td>
<td><strong>0.16</strong></td>
<td><strong>0.27</strong></td>
<td><strong>0.56</strong></td>
<td><strong>0.53</strong></td>
<td><strong>0.33</strong></td>
<td><strong>0.30</strong></td>
<td><strong>0.59</strong></td>
</tr>
<tr>
<td>F1 score Single</td>
<td></td>
<td><strong>U-HFD</strong></td>
<td><strong>0.45</strong></td>
<td><strong>0.09</strong></td>
<td><strong>0.65</strong></td>
<td><strong>0.92</strong></td>
<td><strong>0.04</strong></td>
<td><strong>0.10</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.81</strong></td>
<td><strong>0.09</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>U-LH-2.0</strong></td>
<td><strong>0.23</strong></td>
<td><strong>0.07</strong></td>
<td><strong>0.23</strong></td>
<td><strong>0.29</strong></td>
<td><strong>0.05</strong></td>
<td><strong>0.06</strong></td>
<td><strong>0.21</strong></td>
<td><strong>0.28</strong></td>
<td><strong>0.05</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>U-LH-1.4</strong></td>
<td><strong>0.23</strong></td>
<td><strong>0.09</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.40</strong></td>
<td><strong>0.00</strong></td>
<td><strong>0.07</strong></td>
<td><strong>0.31</strong></td>
<td><strong>0.35</strong></td>
<td><strong>0.06</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>ACL</strong></td>
<td><strong>0.23</strong></td>
<td><strong>0.07</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.25</strong></td>
<td><strong>0.04</strong></td>
<td><strong>0.05</strong></td>
<td><strong>0.17</strong></td>
<td><strong>0.20</strong></td>
<td><strong>0.04</strong></td>
</tr>
<tr>
<td>F1 score Multiple</td>
<td></td>
<td><strong>U-HFD</strong></td>
<td><strong>0.59</strong></td>
<td><strong>0.50</strong></td>
<td><strong>0.69</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.19</strong></td>
<td><strong>0.36</strong></td>
<td><strong>0.91</strong></td>
<td><strong>0.89</strong></td>
<td><strong>0.33</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>U-LH-2.0</strong></td>
<td><strong>0.59</strong></td>
<td><strong>0.42</strong></td>
<td><strong>0.73</strong></td>
<td><strong>0.77</strong></td>
<td><strong>0.22</strong></td>
<td><strong>0.25</strong></td>
<td><strong>0.65</strong></td>
<td><strong>0.62</strong></td>
<td><strong>0.17</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>U-LH-1.4</strong></td>
<td><strong>0.52</strong></td>
<td><strong>0.45</strong></td>
<td><strong>0.73</strong></td>
<td><strong>0.90</strong></td>
<td><strong>0.27</strong></td>
<td><strong>0.29</strong></td>
<td><strong>0.79</strong></td>
<td><strong>0.77</strong></td>
<td><strong>0.20</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>ACL</strong></td>
<td><strong>0.59</strong></td>
<td><strong>0.25</strong></td>
<td><strong>0.70</strong></td>
<td><strong>0.64</strong></td>
<td><strong>0.20</strong></td>
<td><strong>0.19</strong></td>
<td><strong>0.51</strong></td>
<td><strong>0.49</strong></td>
<td><strong>0.14</strong></td>
</tr>
</tbody>
</table>
Local clustering on a hypergraph constructed from Microsoft academic coauthorship data

<table>
<thead>
<tr>
<th>Metric</th>
<th>Method</th>
<th>Data</th>
<th>ML</th>
<th>TCS</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond</td>
<td>U-HFD</td>
<td>0.03</td>
<td>0.06</td>
<td>0.06</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>U-LH-2.0</td>
<td>0.07</td>
<td>0.09</td>
<td>0.10</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>U-LH-1.4</td>
<td>0.07</td>
<td>0.08</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>ACL</td>
<td>0.08</td>
<td>0.11</td>
<td>0.11</td>
<td>0.09</td>
</tr>
<tr>
<td>F1 score</td>
<td>U-HFD</td>
<td>0.78</td>
<td>0.54</td>
<td>0.86</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>U-LH-2.0</td>
<td>0.67</td>
<td>0.46</td>
<td>0.71</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>U-LH-1.4</td>
<td>0.65</td>
<td>0.46</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>ACL</td>
<td>0.64</td>
<td>0.43</td>
<td>0.70</td>
<td>0.57</td>
</tr>
</tbody>
</table>

**Nodes** are papers
**Hyperedges** are papers having at least a common coauthor
**Clusters** are papers published at similar venues
Hyper-Flow Diffusion: empirical results

Local clustering on a hypergraph constructed from travel metasearch data (F1 scores)

<table>
<thead>
<tr>
<th>Method</th>
<th>South Korea</th>
<th>Iceland</th>
<th>Puerto Rico</th>
<th>Crimea</th>
<th>Vietnam</th>
<th>Hong Kong</th>
<th>Malta</th>
<th>Guatemala</th>
<th>Ukraine</th>
<th>Estonia</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-HFD</td>
<td>0.75</td>
<td>0.99</td>
<td>0.89</td>
<td>0.85</td>
<td>0.28</td>
<td>0.82</td>
<td>0.98</td>
<td>0.94</td>
<td>0.60</td>
<td>0.94</td>
</tr>
<tr>
<td>C-HFD</td>
<td>0.76</td>
<td>0.99</td>
<td>0.95</td>
<td>0.94</td>
<td>0.32</td>
<td>0.80</td>
<td>0.98</td>
<td>0.97</td>
<td>0.68</td>
<td>0.94</td>
</tr>
<tr>
<td>U-LH-2.0</td>
<td>0.70</td>
<td>0.86</td>
<td>0.79</td>
<td>0.70</td>
<td>0.24</td>
<td>0.92</td>
<td>0.88</td>
<td>0.82</td>
<td>0.50</td>
<td>0.90</td>
</tr>
<tr>
<td>C-LH-2.0</td>
<td>0.73</td>
<td>0.90</td>
<td>0.84</td>
<td>0.78</td>
<td>0.27</td>
<td><strong>0.94</strong></td>
<td>0.96</td>
<td>0.88</td>
<td>0.51</td>
<td>0.83</td>
</tr>
<tr>
<td>U-LH-1.4</td>
<td>0.69</td>
<td>0.84</td>
<td>0.80</td>
<td>0.75</td>
<td>0.28</td>
<td>0.87</td>
<td>0.92</td>
<td>0.83</td>
<td>0.47</td>
<td>0.90</td>
</tr>
<tr>
<td>C-LH-1.4</td>
<td>0.71</td>
<td>0.88</td>
<td>0.84</td>
<td>0.78</td>
<td>0.27</td>
<td>0.88</td>
<td>0.93</td>
<td>0.85</td>
<td>0.50</td>
<td>0.85</td>
</tr>
<tr>
<td>ACL</td>
<td>0.65</td>
<td>0.84</td>
<td>0.75</td>
<td>0.68</td>
<td>0.23</td>
<td>0.90</td>
<td>0.83</td>
<td>0.69</td>
<td>0.50</td>
<td>0.88</td>
</tr>
</tbody>
</table>

**Nodes** are hotel accommodations

**Hyperedges** are accommodations viewed by the same user in a browsing session

**Clusters** are accommodations located in the same country/territory
Hyper-Flow Diffusion: empirical results

Node-ranking and local clustering results on a Florida Bay food network.

<table>
<thead>
<tr>
<th>Method</th>
<th>Query: Raptors</th>
<th>Query: Gray Snapper</th>
<th>Clustering F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Epiphytic Gastropods, Detriti. Gastropods</td>
<td>Meiofauna, Epiphytic Gastropods</td>
<td>Prod.</td>
</tr>
<tr>
<td>U-HFD</td>
<td></td>
<td></td>
<td>0.69</td>
</tr>
<tr>
<td>C-HFD</td>
<td></td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>S-HFD</td>
<td>Gruiformes, Small Shorebirds</td>
<td>Snook, Mackerel</td>
<td>0.69</td>
</tr>
</tbody>
</table>

- S-HFD uses specialized submodular cut-cost shown on the left.
- The example shows that general submodular cut-cost can be necessary.
- HFD is the only local diffusion method that works with general submodular cut-costs.
Hyper-Flow Diffusion: empirical results

For more experiments and details on both synthetic and real datasets:

Please see our preprint Local Hyper-Flow Diffusion on arXiv:2102.07945

Julia implementation HFD on GitHub
Thank you!