Structural sparsity in the real world

Erik Demaine*, Felix Reidl, Peter Rossmanith, Fernando Sánchez Villaamil, Blair D. Sullivan† and Somnath Sikdar

Theoretical Computer Science

MIT †NCSU

@Bergen 2015
Contents

The Program

Structural Sparseness

Models

Algorithms

Empirical Sparseness
The Program
Complex networks

Ubiquitous in real world

Empirical structure
  - Small-world
  - Heavy-tailed degree seq.
  - Clustering

Algorithmic applications
  - Disease spreading
  - Attack resilience
  - Fraud detection
  - Drug discovery

Structural graph theory

Well-researched

Deep structural theorems
  - WQO by minor relation
  - Decomposition theorems
  - Grid-theorem

Great algorithmic properties
  - (E)PTAS
  - Subexponential algorithms
  - Linear kernels
  - Model-checking

Can we bring these two fields together?
The idea

1. **Bridge the gap** by identifying a notion of sparseness that applies to complex networks.
2. **Develop** algorithmic tools for network related problems.
3. **Show experimentally** that the above is useful in practice.
The idea

1 Bridge the gap by identifying a notion of sparseness that applies to complex networks.
   - Need general and stable notion of sparseness.
   - How to prove that it holds for complex networks?

2 Develop algorithmic tools for network related problems.
   - Unclear what problems are interesting.

3 Show experimentally that the above is useful in practice.
   - Show that structural sparseness appears in the real world.
   - Show that algorithms can compete with known approaches.
Structural Sparseness
Star forests
Bounded treedepth
Bounded treewidth
Excluding a minor
Excluding a topological minor
Bounded expansion
Outerplanar
Planar
Bounded genus
Bounded degree
Locally excluding a minor
Locally bounded treewidth
Locally bounded expansion
Nowhere dense
Forests
Star forests
Linear forests
Bounded degree
Bounded expansion
Excluding a minor
Bounded expansion

A graph class has **bounded expansion** if the density of its minors only depends on their **depth**.

The following operations on a class of bounded expansion result again in a class of bounded expansion:

- Taking **shallow minors/immersions** (in particular subgraphs)
- Adding a **universal vertex**
- Replacing each vertex by a **small clique** (lexicographic product)
Models
Perturbed bounded degree

Stochastic Block

Kleinberg

Configuration

Chung-Lu

Barabasi-Albert

Heavy-tailed degree distribution
The positive side

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition $f(d)$</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power law</td>
<td>$d^{-\gamma}$</td>
<td>$\gamma &gt; 2$</td>
</tr>
<tr>
<td>Power law w/ cutoff</td>
<td>$d^{-\gamma} e^{-\lambda d}$</td>
<td>$\gamma &gt; 2$, $\lambda &gt; 0$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$e^{-\lambda d}$</td>
<td>$\lambda &gt; 0$</td>
</tr>
<tr>
<td>Stretched exponential</td>
<td>$d^{\beta-1} e^{-\lambda d^\beta}$</td>
<td>$\lambda, \beta &gt; 0$</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$\exp\left(-\frac{(d-\mu)^2}{2\sigma^2}\right)$</td>
<td>$\mu, \sigma$</td>
</tr>
<tr>
<td>Log-normal</td>
<td>$d^{-1} \exp\left(-\frac{(\log d - \mu)^2}{2\sigma^2}\right)$</td>
<td>$\mu, \sigma$</td>
</tr>
</tbody>
</table>

**Theorem**

Let $\mathcal{D}$ be an asymptotic degree distribution with finite mean. Then random graphs generated by the Configuration Model or the Chung-Lu model with parameter $\mathcal{D}$ have bounded expansion with high probability.
Theorem
The perturbed bounded degree model has bounded expansion with high probability.

Perturbing forests of $S_{\sqrt{n}}$ results in a somewhere dense class.
The negative side

Theorem
The Kleinberg Model is somewhere dense with high probability.

Theorem
The Barabási-Albert Model is somewhere dense with non-vanishing probability.
Bounded expansion

Perturbed bounded degree

Stochastic Block

Kleinberg

Somewhere dense

Heavy-tailed degree distribution

Configuration

Chung-Lu

Barabasi-Albert

Chung-Lu

4

E[d]

3

1

4

\(\Pi(k) \propto k\)
Algorithms
## Neighbourhood sizes

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
<th>Localized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closeness</td>
<td>$\left( \sum_{u \in V(G)} d(v, u) \right)^{-1}$</td>
<td>$\left( \sum_{u \in N^r(v)} d(v, u) \right)^{-1}$</td>
</tr>
<tr>
<td>Harmonic</td>
<td>$\sum_{u \in V(G)} d(v, u)^{-1}$</td>
<td>$\sum_{u \in N^r(v)} d(v, u)^{-1}$</td>
</tr>
<tr>
<td>Lin’s index</td>
<td>$\frac{\left</td>
<td>{ v \mid d(v, v) &lt; \infty } \right</td>
</tr>
</tbody>
</table>

**Theorem**

Let $\mathcal{G}$ be a graph class of bounded expansion. There is an algorithm that for every $r \in \mathbb{N}$ and $G \in \mathcal{G}$ computes the size of the $i$-th neighbourhood of every vertex of $G$, for all $i \leq r$, in linear time.
Closeness centrality

\[
\left( \sum_{u \in N^1(v)} d(v, u) \right)^{-1}
\]
Closeness centrality

\[
\left( \sum_{u \in N^2(v)} d(v, u) \right)^{-1}
\]
Closeness centrality

\[
\left( \sum_{u \in N^3(v)} d(v, u) \right)^{-1}
\]

Network provided by Pål
Closeness centrality

\[
\left( \sum_{u \in N^4(v)} d(v, u) \right)^{-1}
\]
Top-10% recovery

Jaccard similarity of top 10%

Percentage of diameter

Netscience
Codeminer
Diseasome
Cpan-distr.
HepTh
CondMat
Theorem
Given a graph $H$ on $h$ vertices, a graph $G$ on $n$ vertices and a treedepth decomposition of $G$ of height $t$, one can compute the
- number of isomorphisms from $H$ to subgraphs of $G$,
- homomorphisms from $H$ to subgraphs of $G$, or
- (induced) subgraphs of $G$ isomorphic to $H$
in time $O(8^h \cdot t^h \cdot h^2 \cdot n)$ and space $O(4^h \cdot t^h \cdot ht \cdot \log n)$. 
Counting substructures

Theorem (Nešetřil & Ossona de Mendez)
Let \( \mathcal{G} \) be class of bounded expansion. There exists a function \( f \) such that for every \( p \), every member of \( \mathcal{G} \) has a \( p \)-centered coloring with at most \( f(p) \) colors. Moreover, such a coloring can be computed in linear time.
Counting substructures

Theorem (Nešetřil & Ossona de Mendez)

Let $\mathcal{G}$ be class of bounded expansion. There exists a function $f$ such that for every $p$, every member of $\mathcal{G}$ has a $p$-centered coloring with at most $f(p)$ colors. Moreover, such a coloring can be computed in linear time.
5-centered coloring of gcc of netscience graph.
5-centered coloring of gcc of netscience graph.
5-centered coloring of gcc of netscience graph.
How many in a?
Example: Counting $P_4$s

Preprocessing: create $k$-Patterns (here: $k = 2$)

- Take pattern graph $P_4$
- Choose separator
- Choose component
- Label separator
Example: Counting $P_4$s
Example: Counting $P_4$s
Example: Counting $P_4$s
Example: Counting $P_4$s
Example: Counting $P_4$s
Example: Counting $P_4$s
Example: Counting $P_4$s
Example: Counting $P_4$s
Example: Counting $P_4$s
Example: Counting $P_4$s
Example: Counting $P_4$s
Example: Counting $P_4$s
Example: Counting $P_4$s

There are seven $P_4$s in the target graph.
Empirical Sparseness
Closing the gap

In order to claim that our approach is useful in practice we cannot just rely on theory.

- **Graph classes vs. concrete instances**
- The bounds given by our proofs are enormous.
- Random graph models capture only some aspectes of complex networks.
- We prove asymptotic bounds.

(although we show fast convergence)
<table>
<thead>
<tr>
<th>Network</th>
<th>Vertices</th>
<th>Edges</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th></th>
<th>∞</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airlines</td>
<td>235</td>
<td>1297</td>
<td>11</td>
<td>28</td>
<td>39</td>
<td>47</td>
<td>55</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>C.Elegans</td>
<td>306</td>
<td>2148</td>
<td>8</td>
<td>36</td>
<td>74</td>
<td>83</td>
<td>118</td>
<td>153</td>
<td></td>
</tr>
<tr>
<td>Codeminer</td>
<td>724</td>
<td>1017</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>17</td>
<td>23</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>Cpan-authors</td>
<td>839</td>
<td>2212</td>
<td>9</td>
<td>24</td>
<td>34</td>
<td>43</td>
<td>47</td>
<td>224</td>
<td></td>
</tr>
<tr>
<td>Diseasome</td>
<td>1419</td>
<td>2738</td>
<td>12</td>
<td>17</td>
<td>22</td>
<td>25</td>
<td>30</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Polblogs</td>
<td>1491</td>
<td>16715</td>
<td>30</td>
<td>118</td>
<td>286</td>
<td>354</td>
<td>392</td>
<td>603</td>
<td></td>
</tr>
<tr>
<td>Netscience</td>
<td>1589</td>
<td>2742</td>
<td>20</td>
<td>20</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Drosophila</td>
<td>1781</td>
<td>8911</td>
<td>12</td>
<td>65</td>
<td>137</td>
<td>188</td>
<td>263</td>
<td>395</td>
<td></td>
</tr>
<tr>
<td>Yeast</td>
<td>2284</td>
<td>6646</td>
<td>12</td>
<td>38</td>
<td>178</td>
<td>254</td>
<td>431</td>
<td>408</td>
<td></td>
</tr>
<tr>
<td>Cpan-distr.</td>
<td>2719</td>
<td>5016</td>
<td>5</td>
<td>14</td>
<td>32</td>
<td>42</td>
<td>56</td>
<td>224</td>
<td></td>
</tr>
<tr>
<td>Twittercrawl</td>
<td>3656</td>
<td>154824</td>
<td>89</td>
<td>561</td>
<td>1206</td>
<td>1285</td>
<td>1341</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td>4941</td>
<td>6594</td>
<td>6</td>
<td>12</td>
<td>20</td>
<td>21</td>
<td>34</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>AS Jan 2000</td>
<td>6474</td>
<td>13895</td>
<td>12</td>
<td>29</td>
<td>70</td>
<td>102</td>
<td>151</td>
<td>357</td>
<td></td>
</tr>
<tr>
<td>Hep-th</td>
<td>7610</td>
<td>15751</td>
<td>24</td>
<td>25</td>
<td>104</td>
<td>328</td>
<td>360</td>
<td>558</td>
<td></td>
</tr>
<tr>
<td>Gnutella04</td>
<td>10876</td>
<td>39994</td>
<td>8</td>
<td>43</td>
<td>626</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ca-HepPh</td>
<td>12008</td>
<td>118489</td>
<td>239</td>
<td>296</td>
<td>1002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CondMat</td>
<td>16264</td>
<td>47594</td>
<td>18</td>
<td>47</td>
<td>255</td>
<td>1839</td>
<td></td>
<td>1310</td>
<td></td>
</tr>
<tr>
<td>ca-CondMat</td>
<td>23133</td>
<td>93497</td>
<td>26</td>
<td>89</td>
<td>665</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enron</td>
<td>36692</td>
<td>183831</td>
<td>27</td>
<td>214</td>
<td>1428</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brightkite</td>
<td>58228</td>
<td>214078</td>
<td>39</td>
<td>193</td>
<td>1421</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network</td>
<td>Vertices</td>
<td>Edges</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>----------</td>
<td>--------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>Airlines</td>
<td>235</td>
<td>1297</td>
<td>11</td>
<td>28</td>
<td>39</td>
<td>47</td>
<td>55</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td>4941</td>
<td>6594</td>
<td>6</td>
<td>12</td>
<td>20</td>
<td>21</td>
<td>34</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>AS Jan 2000</td>
<td>6474</td>
<td>13895</td>
<td>12</td>
<td>29</td>
<td>70</td>
<td>102</td>
<td>151</td>
<td>357</td>
<td></td>
</tr>
<tr>
<td>C.Elegans</td>
<td>306</td>
<td>2148</td>
<td>8</td>
<td>36</td>
<td>74</td>
<td>83</td>
<td>118</td>
<td>153</td>
<td></td>
</tr>
<tr>
<td>Diseaseome</td>
<td>1419</td>
<td>2738</td>
<td>12</td>
<td>17</td>
<td>22</td>
<td>25</td>
<td>30</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Drosophila</td>
<td>1781</td>
<td>8911</td>
<td>12</td>
<td>65</td>
<td>137</td>
<td>188</td>
<td>263</td>
<td>395</td>
<td></td>
</tr>
<tr>
<td>Yeast</td>
<td>2284</td>
<td>6646</td>
<td>12</td>
<td>38</td>
<td>178</td>
<td>254</td>
<td>431</td>
<td>408</td>
<td></td>
</tr>
<tr>
<td>Codeminer</td>
<td>724</td>
<td>1017</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>17</td>
<td>23</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>Gnutella04</td>
<td>10876</td>
<td>39994</td>
<td>8</td>
<td>43</td>
<td>626</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Enron</td>
<td>36692</td>
<td>183831</td>
<td>27</td>
<td>214</td>
<td>1428</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Brightkite</td>
<td>58228</td>
<td>214078</td>
<td>39</td>
<td>193</td>
<td>1421</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Cpan-authors</td>
<td>839</td>
<td>2212</td>
<td>9</td>
<td>24</td>
<td>34</td>
<td>43</td>
<td>47</td>
<td>224</td>
<td></td>
</tr>
<tr>
<td>Polblogs</td>
<td>1491</td>
<td>16715</td>
<td>30</td>
<td>118</td>
<td>286</td>
<td>354</td>
<td>392</td>
<td>603</td>
<td></td>
</tr>
<tr>
<td>Netscience</td>
<td>1589</td>
<td>2742</td>
<td>20</td>
<td>20</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Cpan-distr.</td>
<td>2719</td>
<td>5016</td>
<td>5</td>
<td>14</td>
<td>32</td>
<td>42</td>
<td>56</td>
<td>224</td>
<td></td>
</tr>
<tr>
<td>Twittercrawl</td>
<td>3656</td>
<td>154824</td>
<td>89</td>
<td>561</td>
<td>1206</td>
<td>1285</td>
<td>1341</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Hep-th</td>
<td>7610</td>
<td>15751</td>
<td>24</td>
<td>25</td>
<td>104</td>
<td>328</td>
<td>360</td>
<td>558</td>
<td></td>
</tr>
<tr>
<td>ca-HepPh</td>
<td>12008</td>
<td>118489</td>
<td>239</td>
<td>296</td>
<td>1002</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>CondMat</td>
<td>16264</td>
<td>47594</td>
<td>18</td>
<td>47</td>
<td>255</td>
<td>1839</td>
<td>–</td>
<td>1310</td>
<td></td>
</tr>
<tr>
<td>ca-CondMat</td>
<td>23133</td>
<td>93497</td>
<td>26</td>
<td>89</td>
<td>665</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Network</td>
<td>Vertices</td>
<td>Edges</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>( \infty )</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>----------</td>
<td>--------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>Airlines</td>
<td>235</td>
<td>1297</td>
<td>1.00</td>
<td>2.55</td>
<td>3.55</td>
<td>4.27</td>
<td>5.00</td>
<td>5.82</td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td>4941</td>
<td>6594</td>
<td>1.00</td>
<td>2.00</td>
<td>3.33</td>
<td>3.50</td>
<td>5.67</td>
<td>15.83</td>
<td></td>
</tr>
<tr>
<td>AS Jan 2000</td>
<td>6474</td>
<td>13895</td>
<td>1.00</td>
<td>2.42</td>
<td>5.83</td>
<td>8.50</td>
<td>12.58</td>
<td>29.75</td>
<td></td>
</tr>
<tr>
<td>C.Elegans</td>
<td>306</td>
<td>2148</td>
<td>1.00</td>
<td>4.50</td>
<td>9.25</td>
<td>10.38</td>
<td>14.75</td>
<td>19.12</td>
<td></td>
</tr>
<tr>
<td>Diseasome</td>
<td>1419</td>
<td>2738</td>
<td>1.00</td>
<td>1.42</td>
<td>1.83</td>
<td>2.08</td>
<td>2.50</td>
<td>2.50</td>
<td></td>
</tr>
<tr>
<td>Drosophila</td>
<td>1781</td>
<td>8911</td>
<td>1.00</td>
<td>5.42</td>
<td>11.42</td>
<td>15.67</td>
<td>21.92</td>
<td>32.92</td>
<td></td>
</tr>
<tr>
<td>Yeast</td>
<td>2284</td>
<td>6646</td>
<td>1.00</td>
<td>3.17</td>
<td>14.83</td>
<td>21.17</td>
<td>35.92</td>
<td>34.00</td>
<td></td>
</tr>
<tr>
<td>Codeminer</td>
<td>724</td>
<td>1017</td>
<td>1.00</td>
<td>2.00</td>
<td>3.00</td>
<td>3.40</td>
<td>4.60</td>
<td>10.20</td>
<td></td>
</tr>
<tr>
<td>Gnutella04</td>
<td>10876</td>
<td>39994</td>
<td>1.00</td>
<td>5.38</td>
<td>78.25</td>
<td>15.00</td>
<td>23.25</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Enron</td>
<td>36692</td>
<td>183831</td>
<td>1.00</td>
<td>7.93</td>
<td>52.89</td>
<td>15.00</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Brightkite</td>
<td>58228</td>
<td>214078</td>
<td>1.00</td>
<td>4.95</td>
<td>36.44</td>
<td>15.00</td>
<td>15.00</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Cpan-authors</td>
<td>839</td>
<td>2212</td>
<td>1.00</td>
<td>2.67</td>
<td>3.78</td>
<td>4.78</td>
<td>5.22</td>
<td>24.89</td>
<td></td>
</tr>
<tr>
<td>Polblogs</td>
<td>1491</td>
<td>16715</td>
<td>1.00</td>
<td>3.93</td>
<td>9.53</td>
<td>11.80</td>
<td>13.07</td>
<td>20.10</td>
<td></td>
</tr>
<tr>
<td>Netscience</td>
<td>1589</td>
<td>2742</td>
<td>1.00</td>
<td>1.00</td>
<td>1.40</td>
<td>1.40</td>
<td>1.40</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Cpan-distr.</td>
<td>2719</td>
<td>5016</td>
<td>1.00</td>
<td>2.80</td>
<td>6.40</td>
<td>8.40</td>
<td>11.20</td>
<td>44.80</td>
<td></td>
</tr>
<tr>
<td>Twittercrawl</td>
<td>3656</td>
<td>154824</td>
<td>1.00</td>
<td>6.30</td>
<td>13.55</td>
<td>14.44</td>
<td>15.07</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Hep-th</td>
<td>7610</td>
<td>15751</td>
<td>1.00</td>
<td>1.04</td>
<td>4.33</td>
<td>13.67</td>
<td>15.00</td>
<td>23.25</td>
<td></td>
</tr>
<tr>
<td>ca-HepPh</td>
<td>12008</td>
<td>118489</td>
<td>1.00</td>
<td>1.24</td>
<td>4.19</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CondMat</td>
<td>16264</td>
<td>47594</td>
<td>1.00</td>
<td>2.61</td>
<td>14.17</td>
<td>102.17</td>
<td>-</td>
<td>72.78</td>
<td></td>
</tr>
<tr>
<td>ca-CondMat</td>
<td>23133</td>
<td>93497</td>
<td>1.00</td>
<td>3.42</td>
<td>25.58</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

• We show that several important models of complex networks have bounded expansion.
• Besides the known algorithms (first-order model checking!) we show that relevant problems can be solved faster by using this fact.
• Our experiments demonstrate that many networks are structurally sparse.
Conclusion

• We show that several important models of complex networks have bounded expansion.
• Besides the known algorithms (first-order model checking!) we show that relevant problems can be solved faster by using this fact.
• Our experiments demonstrate that many networks are structurally sparse.

THANKS!
Questions?