THE THEORY OF WALKS

AN OVERVIEW

1. Why walks?

2. The theory
   - Algebraic combinatorics
   - Number-theory

3. The applications
   - Graph Generation
   - Algorithmics
   - Social sciences
   - Machine Learning
   - Systems Biology
   - Econometry
   - Matrix Computations
   - Differential Calculus
   - Statistical Inference
WHY WALKS?

Walk-based methods in network-analysis / machine learning

Google scholar data

- Largely empirical approaches

What can we possibly learn from walks on graphs? Everything?
WHY WALKS?

Which graphs are the most similar?
WHY WALKS?

Which graphs are the most similar?
1. Why walks?

2. The theory
   - Algebraic combinatorics
   - Number-theory

3. The applications
   - Algorithmics
   - Social sciences
   - Machine Learning
   - Systems Biology
   - Econometry
   - Matrix Computations
   - Differential Calculus
   - Statistical Inference
Walks = words on the edges $e_{ij}$ we impose

$e_{ij} e_{ik} \neq e_{ik} e_{ij}$

$e_{ij} e_{kl} = e_{kl} e_{ij}$

Equivalence classes on words: heaps of simple cycles “hikes”


$\gamma | h . h' \Rightarrow \gamma | h \text{ ou } \gamma | h'$

$h . h' = h' . h \iff V(h) \cap V(h') = \emptyset$
A THEORY OF WALKS?

- Gregory Lawler (1987)

- Factorisation into simple cycles & a simple path
  
  For $w, w'$ cycles
  
  Simple cycles have prime property $\gamma \mid w.w' \Rightarrow \gamma \mid w$ ou $\gamma \mid w'$

A number theory for walks?!
2) How do we build a theory?

- **Rota et al. (1964-1974):** number theory from posets!

  Order integers by divisibility, construct reduced incidence algebra of the poset:
  
  - Zeta function, von Mangoldt function etc.
  - Dirichlet series
  - Relations between the functions

- Poset of hikes ordered by divisibility $h | h' \Rightarrow h \leq h'$

  \[
  P_{\mathcal{H}} := (\mathcal{H}, \leq)
  \]

  \[
  h \cdot h' = h' \cdot h \iff \mathcal{V}(h) \cap \mathcal{V}(h') = \emptyset
  \]

  Semi-commutative reduced incidence algebra gives rise to...
AN EXTENSION OF NUMBER THEORY

<table>
<thead>
<tr>
<th></th>
<th>Hikes</th>
<th>Number theory</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Zeta</strong></td>
<td>$\zeta = S1 = \sum_{h \in \mathcal{H}} h$</td>
<td>$\zeta_R(s) = \sum_{n&gt;0} \frac{1}{n^s}$</td>
</tr>
<tr>
<td></td>
<td>$\zeta = \frac{1}{\det(I-W)}$</td>
<td></td>
</tr>
<tr>
<td><strong>Möbius</strong></td>
<td>$\mu(h) = \begin{cases} (-1)^{\Omega(h)}, &amp; h \text{ self-avoiding} \ 0, &amp; \text{otherwise.} \end{cases}$</td>
<td>$\mu(n) = \begin{cases} (-1)^{\Omega(n)}, &amp; n \text{ square-free} \ 0, &amp; \text{otherwise.} \end{cases}$</td>
</tr>
<tr>
<td><strong>Von Mangoldt</strong></td>
<td>$\Lambda(h) = \begin{cases} \ell(p), &amp; h \text{ walk, } p</td>
<td>h \ 0, &amp; \text{otherwise.} \end{cases}$</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{S}\Lambda = \text{Tr} \left[ (I-W)^{-1} \right]$</td>
<td></td>
</tr>
<tr>
<td><strong>Liouville</strong></td>
<td>$\lambda(h) = (-1)^{\Omega(h)}$</td>
<td>$\lambda(h) = (-1)^{\Omega(n)}$</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{S}\lambda = \frac{1}{\text{perm}(I-W)}$</td>
<td></td>
</tr>
</tbody>
</table>

**THEOREM**

Rigorously reduces to number theory on a special graph

### AN EXTENSION OF NUMBER THEORY

<table>
<thead>
<tr>
<th></th>
<th>Hikes</th>
<th>Number Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of divisors</td>
<td>$\zeta^2$</td>
<td>$\zeta_R(s)^2$</td>
</tr>
<tr>
<td>Log Zeta</td>
<td>$\log \zeta = \sum_h \frac{\Lambda(h)}{\ell(h)} h$</td>
<td>$\log \zeta_R(s) = \sum_n \frac{\Lambda(n)}{\log(n)} \frac{1}{n^s}$</td>
</tr>
<tr>
<td>Log-Mangoldt</td>
<td>$\ell(h) = \sum_{h'</td>
<td>h} \Lambda(h')$</td>
</tr>
<tr>
<td>Totally multiplicative</td>
<td>$f^{-1} = \sum_h \mu(h) f(h) h$</td>
<td>$f^{-1} = \sum_n \frac{\mu(n)f(n)}{n^s}$</td>
</tr>
<tr>
<td>functions</td>
<td>$f'/f = -\sum_h \Lambda(h) f(h) h$</td>
<td>$f'/f = \sum_n \frac{\Lambda(n)f(n)}{n^s}$</td>
</tr>
<tr>
<td>Totally additive</td>
<td>$(f \ast \mu)(h) = \begin{cases} f(p), &amp; h \text{ walk, } p</td>
<td>h \ 0, &amp; \text{otherwise.} \end{cases}$</td>
</tr>
<tr>
<td>functions</td>
<td>$\zeta'/\zeta$ from the</td>
<td>$- \sum_p \log p \frac{p^{-s}}{1-p^{-s}}$</td>
</tr>
<tr>
<td>primes</td>
<td>$- \sum_{\gamma: \text{simple cycle}} \ell(\gamma) \frac{\det(1 - W_{\setminus \gamma})}{\det(1 - W)}$</td>
<td></td>
</tr>
<tr>
<td>Number $\Omega$ of prime factors</td>
<td>$\sum_{w: \text{walk}} w = \det(1 - W) \sum_{h \in \mathcal{H}} \Omega(h) \frac{1}{h}$</td>
<td>$\sum_{p,n} \frac{1}{p^{-ns}} = \zeta_R(s)^{-1} \sum_n \frac{\Omega(n)}{n^s}$</td>
</tr>
</tbody>
</table>
AN EXAMPLE OF THE THEORY

\[ P(z) := \sum_{\gamma: \text{cycle simple}} z^{\ell(\gamma)} \]

Count simple cycles?

\[ \mu \circ \Lambda = \frac{d}{dz} P(z) = \sum_{H \prec G} \text{Tr} \left( (zA_H)^{|H|} (I - zA_H)^{|N(H)|} \right) \]

\[ \Lambda \circ \mu = \frac{d}{dz} P(z) = \sum_{H \prec G} \det(-zA_H) \frac{d}{dz} \text{perm}(I + zA_{G-H}) \]

1. Why walks?

2. The theory

   - Algebraic combinatorics
   - Number-theory

3. The applications

   - Graph Generation
   - Algorithmics
   - Social sciences
   - Machine Learning
   - Systems Biology
   - Econometry
   - Matrix Computations
   - Differential Calculus
   - Statistical Inference
**SO WHAT?**

*Direct consequences: Graph Generation*

- **Problem:** generate co-spectral pairs

- **Solution:** transformation of $G$ preserving poset $P^H$

---

Same spectrum, same immanants, same Weisfeiler-Lehman colours & more!

---

SO WHAT?

Direct consequences: *Algorithmics*

- Problem: counting simple cycles / paths
  - All of them: \#P-complete
  - Up to length $\ell$: \#W[1]-complete

What about our formula?

$$\frac{d}{dz} P(z) = \sum_{H \subset G \text{ connected}} \text{Tr} \left( (zA_H)^{|H|} (1 - zA_H)^{|N(H)|} \right)$$

Up to length $\ell$:

Requires *small* connected subgraphs (reverse search) $|H| \leq \ell$
SO WHAT?

Direct consequences: Algorithmics

‣ Algorithm: counting until $\ell$ costs $O(V + E + (\ell^\omega + \ell\Delta)|S_\ell|)$

Best general purpose algorithm if $(1 + \ell^{\omega-1}/\Delta)|S_\ell| \leq \pi(\ell)$

"Less connected induced subgraphs than simple cycles"
...otherwise brute force search wins

‣ Unfortunately is it an open problem to determine when $|S_\ell| \leq \pi(\ell)$

‣ Sparse graphs $E = O(V)$ algorithm is linear in $V$

Matlab File Exchange « CycleCount, PathCount, CyclePathCount »
Direct consequences: Network Analysis

Nodes \( \rightarrow \) People
Edges \( \rightarrow \) Relations
Sign \( \leftrightarrow +/- = \) Amity/Enmity

Conjecture [Heider 1946]: Balanced cycles should be largely predominant

Verifying Heider’s conjecture is NP-hard

Idea: let’s use our algorithm for weighted simple cycles!
Matlab File Exchange « CycleCount, PathCount, CyclePathCount »
SO WHAT?

Direct consequences: Network Analysis

Answer: Heider was right ... up to a point

**SO WHAT?**

*Direct consequences: Machine Learning*

- **Problem:** automatic graph classification

Graph $A \leftrightarrow$ set $X_A$ of objects

Kernel "a measure of similarity"

$$K(A, B) = \sum_{x_i \in X_A} \sum_{x_j \in X_B} k(x_i, x_j)$$

Base kernel, compares individual objects

---

SO WHAT?

Direct consequences: Machine Learning

- Borgwardt, Kriegel (2005):
  Ideally, set $X_A$ of objects = set of simple paths “all-paths” kernel
  “Computing the all-paths kernel is NP-hard”

Look only at shortest paths “shortest-path kernel”

- Idea: let’s use our algorithms for labeled paths!

Matlab File Exchange « CycleCount, PathCount, CyclePathCount »
SO WHAT?

Direct consequences: **Machine Learning**

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MUTAG</td>
</tr>
<tr>
<td>AllPaths</td>
<td>89.8</td>
</tr>
<tr>
<td>ShortestPath</td>
<td>87.1</td>
</tr>
<tr>
<td>Graphlets</td>
<td>85.2</td>
</tr>
<tr>
<td>WL</td>
<td>88.8</td>
</tr>
<tr>
<td>WL-OA</td>
<td>82.8</td>
</tr>
</tbody>
</table>

Table 1: Classification accuracies (%) on standard graph datasets.

Giscard, Wilson, "The All-paths Graph Kernel", to appear on the arXiv within a week or two (2017)

SO WHAT?

Direct consequences: **Machine Learning**

<table>
<thead>
<tr>
<th>Kernel</th>
<th>MUTAG</th>
<th>PTC-MR</th>
<th>NCI1</th>
<th>NCI109</th>
<th>ENZYMES</th>
</tr>
</thead>
<tbody>
<tr>
<td>AllPaths</td>
<td>2.6</td>
<td>-</td>
<td>36.4</td>
<td><strong>35.7</strong></td>
<td>14.3</td>
</tr>
<tr>
<td>ShortestPath</td>
<td>1.5</td>
<td>-</td>
<td>262</td>
<td>296</td>
<td>6.7</td>
</tr>
<tr>
<td>WL</td>
<td>0.29</td>
<td>-</td>
<td><strong>9.8</strong></td>
<td>-</td>
<td><strong>1.7</strong></td>
</tr>
<tr>
<td>WL-OA</td>
<td>0.34</td>
<td>-</td>
<td>1581</td>
<td>-</td>
<td>89.9</td>
</tr>
</tbody>
</table>

Table 1: Computation time (in seconds) on standard graph datasets.

---

Giscard, Wilson, “The All-paths Graph Kernel”, to appear on the arXiv within a week or two (2017)

Time scales linearly with graph size!!
SO WHAT?

Direct consequences: *Systems biology!*

- Problem: which proteins are targeted by pathogens? why?

*Arabidopsis thaliana*

*Hyaloperonospora arabidopsidis*
SO WHAT?

Direct consequences: **Systems biology**!

- **Model**: targets are hubs *(Mukhthar et al. 2011, Science)*

> “protein targeting by pathogens cannot be explained merely by the high connectivity of those target [proteins]”
SO WHAT?

Direct consequences: Systems biology!

- **Facts:** Pathogens stimulate immune activity
  Central proteins involved in immune interactions

- **Idea:** Triads “Target - Central - Immune” (TCI) exist & targeted triads maximally disrupt flows of proteins reactions

Flow of protein interactions passing through $\forall = \text{number of walks multiples of } \forall$

Non-commutative Brun sieve!

\[
f = \det \left( I - \frac{1}{\lambda} A_{G\setminus \forall} \right)
\]

- Computationally cheap
- Numerically well cond. $0 \leq f \leq 1$
- Induces the eigenvector centrality!

- **Model:** triads with targets are TCI and $f$- dominant

SO WHAT?

Direct consequences: Systems biology!

- It works!!

“pathogens aim at disrupting the largest possible fraction of sequences of protein interactions”

Bonus: we identified 2 proteins surrounded by 70% of all targets

SO WHAT?

Direct consequences: *Econometry*

- **Problem:** evaluate the impact events on capital flow intercepted by chosen ensemble $\gamma$ of economic sectors

- **Solution:** fraction of capital flow is $f = \det \left( \mathbf{I} - \frac{1}{\lambda} \mathbf{A}_{G \setminus \gamma} \right)$

---

SO WHAT?

Direct consequences: Matrix Computations

Problem: \( M \) a matrix, compute \( f(M) \)

Observe

\[
(M^n)_{ij} = \sum_{\substack{w: i \leftrightarrow j \\ell(w)=n}} \text{weight}(w)
\]

\[
f(M) = \sum_n f_n M^n \quad \text{is a series over all walks}
\]

Idea: all walks are multiples of a few primes

Sum only over these primes, generate their multiples via exact formulas

It works! All \( f(M) \) are finite continued fractions over the primes!

SO WHAT?

Direct consequences: *Differential Calculus*

- **Problem:** $M(t)$ a time-dependent matrix, compute $U(t, t')$ such that
  \[ M(t)U(t, t') = \frac{d}{dt}U(t, t') \]

- Central to questions in *quantum dynamics* and *NMR*

- $U(t, t')$ admits a closed expression in terms of prime walks!

  Can solve all systems of linear differential equations

  Extends to partial & fractional diff. equation!

**SO WHAT?**

Direct consequences: *Statistical Inference*

- **Data:** a vector $X$ of random variables, some correlated known from *noisy* observations $Y$

- **Problem:** Compute the marginals of $X|Y$

Marginals of $X|Y$ admit a closed expression in terms of prime walks on the GMRF!

GLOBAL VISION

Theory of Walks

- Bialgeras & idempotents
- Algebraic combinatorics
- Trace monoids
- Cospectral partners
- Machine learning
  - Graph kernel
- Exact formulas
- Connective constants
- Algorithmics & Computations
- Quantum dynamics
- NMR
- Network analysis
- Social sciences
  - Econometry
  - Systems biology
- Number theory
  - Sieves

A SLOW AND SYSTEMATIC APPROACH
THANK YOU!
OUTLINE

1. Why walks?

2. The theory

3. The applications

4. Into the rabbit hole
Asymptotic growth on both sides of this equation is unknown!
CONNECTIVE CONSTANTS

« A widely open problem » Flajolet & Sedgewick

How many self-avoiding polygons?

Asymptotics $\ell \to \infty$?

$\mu \ell^{11/32}$

Rigorously the semi-commutative extension of the prime number theorem

- Non-commutative sieves Brun, Selberg
- Algebraic idempotents (L-functions, $\zeta$, $\Psi$, …)
- Abelianisation (Hochschild homology of $\mathcal{H}$)