Maximal Entropy Random Walk

the most random among random walks
(maximizing entropy production)

RW for minimal information about a system
in agreement with the maximum entropy principle.

strong localization property, **scale-free, nonlocal**

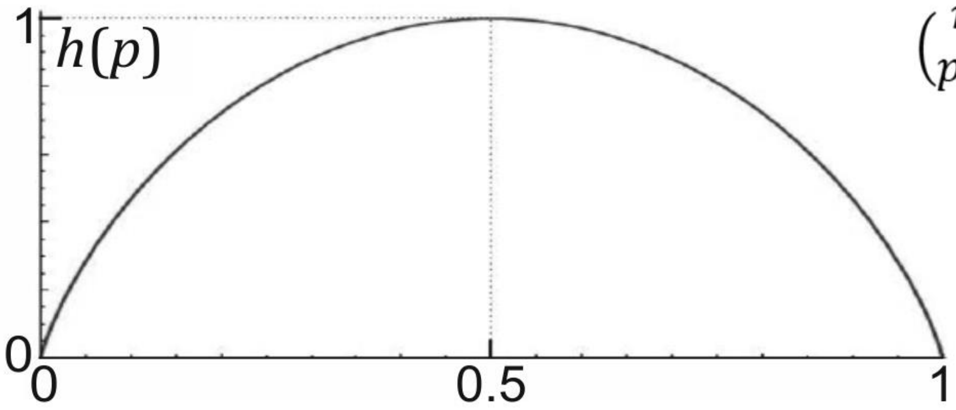
Some applications:

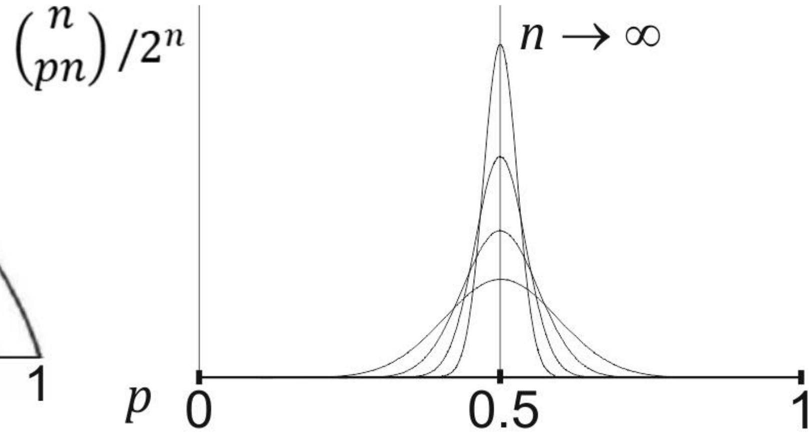
- **maximizing informational capacity** of channel under some constraints (data storage/transmission, maybe linguistics (?)),
- corrections to **diffusion models** to get agreement with quantum predictions (diffusion, conductance, molecular dynamics),
- **metrics for complex networks, data mining** (e.g. centrality measure, saliency regions, PageRank, SimRank, community detection)

Jarosław Duda, UJ
Będlewo 19.02.2015

We need n bits of information to choose one of 2^n possibilities.

For length n 0/1 sequences with pn of "1", how many bits we need to choose one?





$$\binom{n}{pn} \approx 2^{nh(p)} \quad \text{for} \quad h(p) := -p\lg(p) - (1-p)\lg(1-p)$$

A sequence of symbols with $(p_s)_{s=0..m-1}$ probability distribution contains asymptotically $\boldsymbol{H = \sum_s p_s \lg(1/p_s)}$ **bits/symbol** ($H \le \lg(m)$)

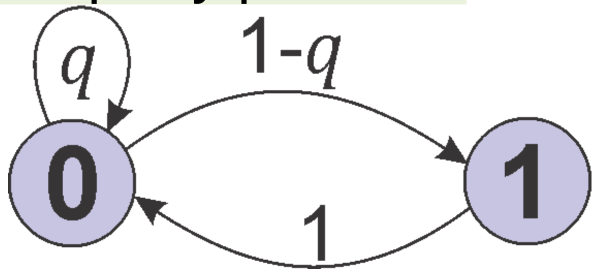
Seen as weighted average:

symbol/event of probability p contains $\lg(1/p)$ bits.

(Jaynes) principle of maximum entropy: while limited knowledge, the safest assumption is probability distribution which maximizes entropy.

Fibonacci coding – as a bit sequence with **constraints**: no two neighboring '1's
e.g. 0010101000010101001001 – each sequence should be equally probable
What about statistics of a single step?

$$M = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \qquad S = \begin{pmatrix} q & 1-q \\ 1 & 0 \end{pmatrix} \qquad q = ?$$



What q should we choose to maximize informational capacity?

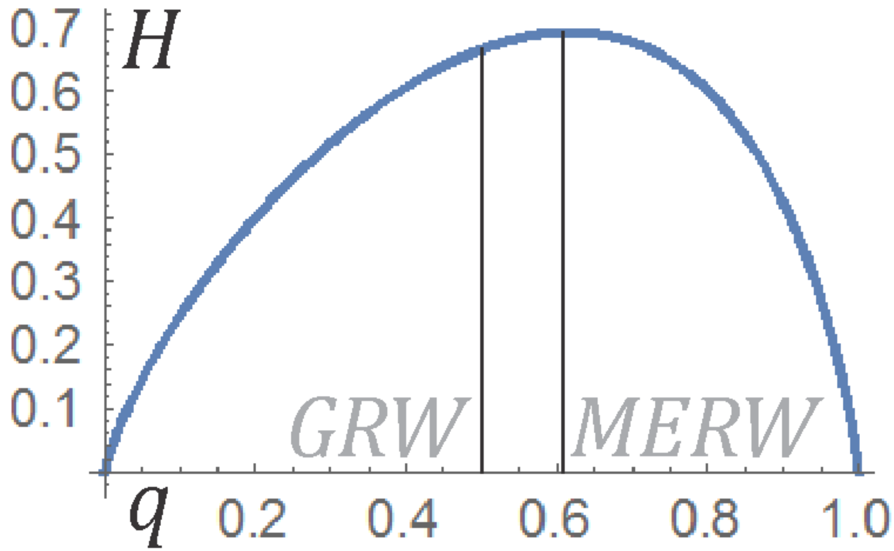
Stationary probability: $\pi = (\Pr(0), \Pr(1))^T$

$$\pi S = \pi$$

$$\pi = \left(\frac{1}{2-q}, 1 - \frac{1}{2-q}\right)$$

Entropy – informational content:

$$H = \sum_i \pi_i \sum_j S_{ij} \lg(1/S_{ij}) = \pi_0 \cdot h(q)$$



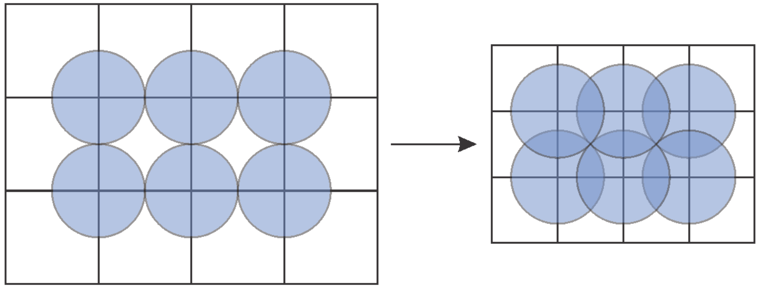
$$H_{max} \approx 0.694241913 \text{ bits/node}$$

$$\text{for} \quad q = \frac{(\sqrt{5}-1)}{2} \approx 0.618034$$

My original MERW motivation: **maximizing capacity under constraints** for 2D analogue of Fibonacci coding ("hard square": no two neighboring '1's)

We get $\boldsymbol{H} \approx \mathbf{0.58789}$ bits/node

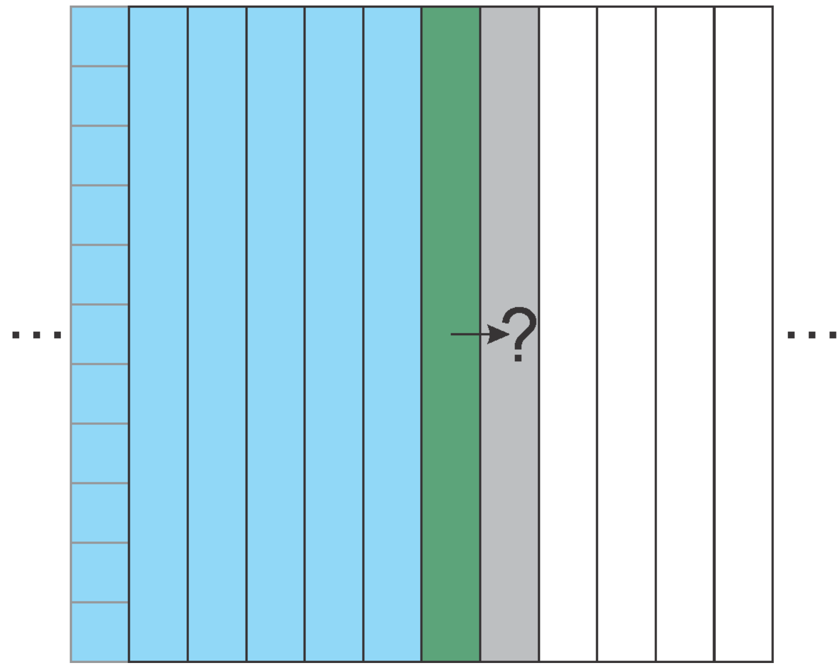
Some application:
use magnetic dots (twice) more densely,
at cost of constraints – two dots cannot overlap.
$2 \cdot 0.58789 \approx 1.\mathbf{176}$



We get 17.6% capacity increase due to better positioning!
(e.g. using 1D MERW on the space of possible succeeding lines)

Approximate with finite width stripe $\infty \times m$
(large) **alphabet**: allowed slices
Adjacency matrix: possible neighbors
… **find MERW** for adjacency matrix… ?

→ translate into local transition probability rules



Graph (M) ❓ **stochastic matrix** (S) $\rightarrow$ **stationary probability** (π)

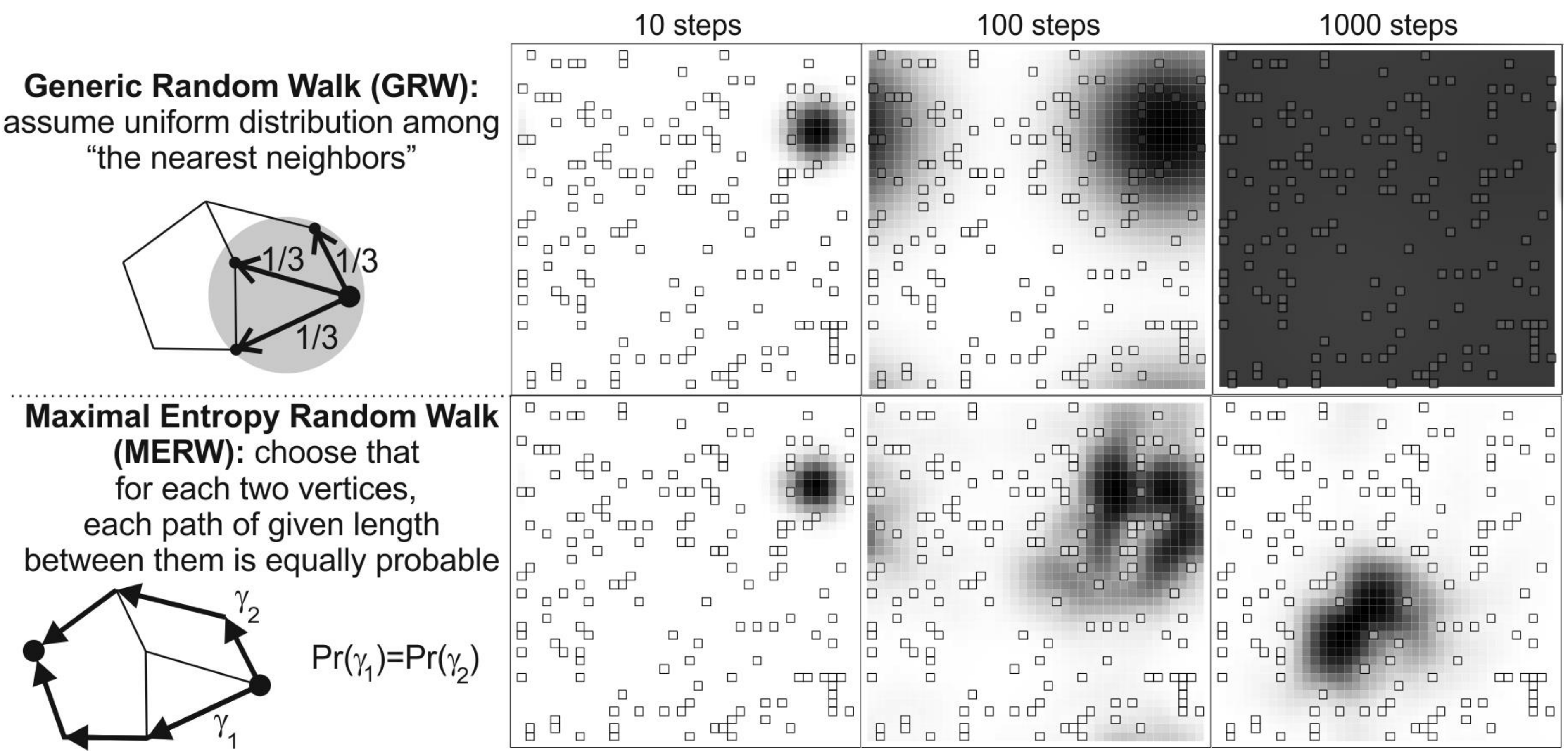
$M_{ab} \in \{0,1\}$ $\quad 0 \le S_{ab} \le M_{ab}$, $\forall_a \sum_b S_{ab} = 1$ $\quad \sum_a \pi_a S_{ab} = \pi_b$

Average **entropy** production per step: $\sum_a \pi_a \sum_b S_{ab} \lg(1/S_{ab})$

GRW and MERW are equal on regular graphs, but e.g. on defected 2D lattice:

Generic Random Walk (GRW): assume uniform distribution among "the nearest neighbors"

Maximal Entropy Random Walk (MERW): choose that for each two vertices, each path of given length between them is equally probable



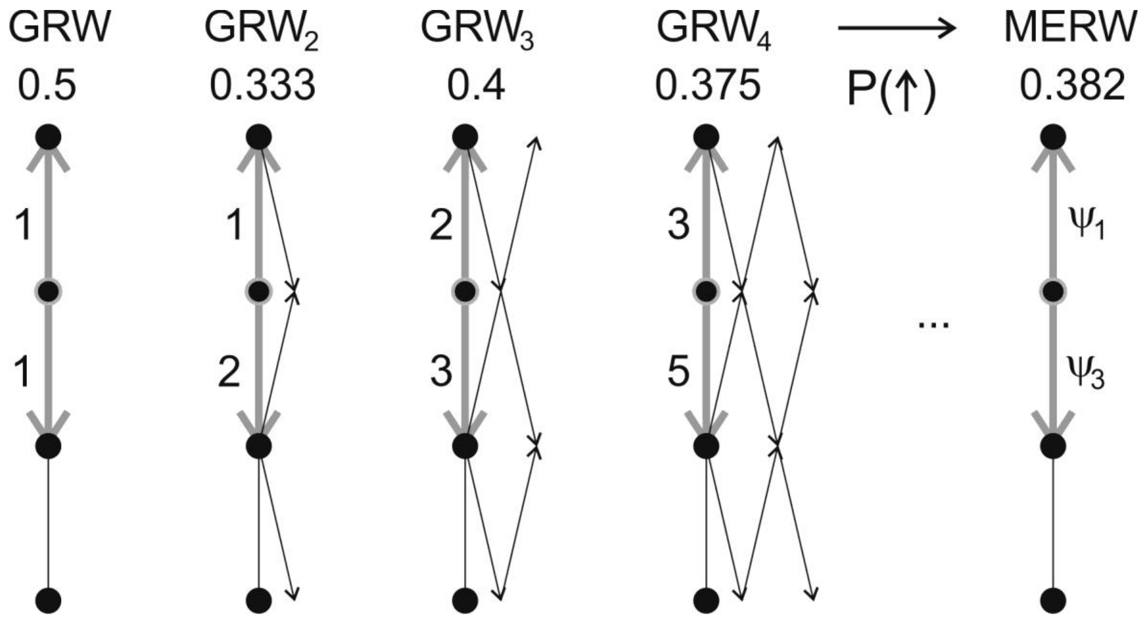
GRW assumes we know exactly the used probabilistic algorithm, has characteristic length

MERW assumes only there are no hidden local probabilistic rules, is scale-free limit of GRW

MERW as scale-free limit of GRW

$$S_{ab}^{GRW_k} \propto M_{ab} \sum_c (M^{k-1})_{bc}$$

GRW: each outgoing **edge** is equally probable $(k = 1)$



GRW_k – each outgoing **length** k **path is equally probable.**

In the limit, the number of paths starting with $a \to b$ **is proportional** to coordinate (ψ_b) of the **dominant eigenvector** of M :

$$M\psi = \lambda\psi$$

Frobenius-Perron theorem for connected graph: real, nondegenerated $\lambda > 0$, $\forall_a\, \psi_a > 0$

Normalization for vertex a: $\sum_b M_{ab}\psi_b = (M\psi)_a = \lambda\psi_a$

Finally: **while being in a, probability of jumping to b is:**
(symmetric M:)

$$\boxed{S_{ab} = \frac{M_{ab}}{\lambda}\frac{\psi_b}{\psi_a}}$$

For which stationary probability distribution $(\pi S = \pi)$ is $\pi_a \propto \psi_a^2$

nonlocality

$$(\pi S)_b = \sum_a \psi_a^2 \cdot \frac{M_{ab}}{\lambda}\frac{\psi_b}{\psi_a} = \sum_a \psi_a M_{ab} \cdot \frac{\psi_b}{\lambda} = \lambda\psi_b \frac{\psi_b}{\lambda} = \psi_b^2 = \pi_b$$

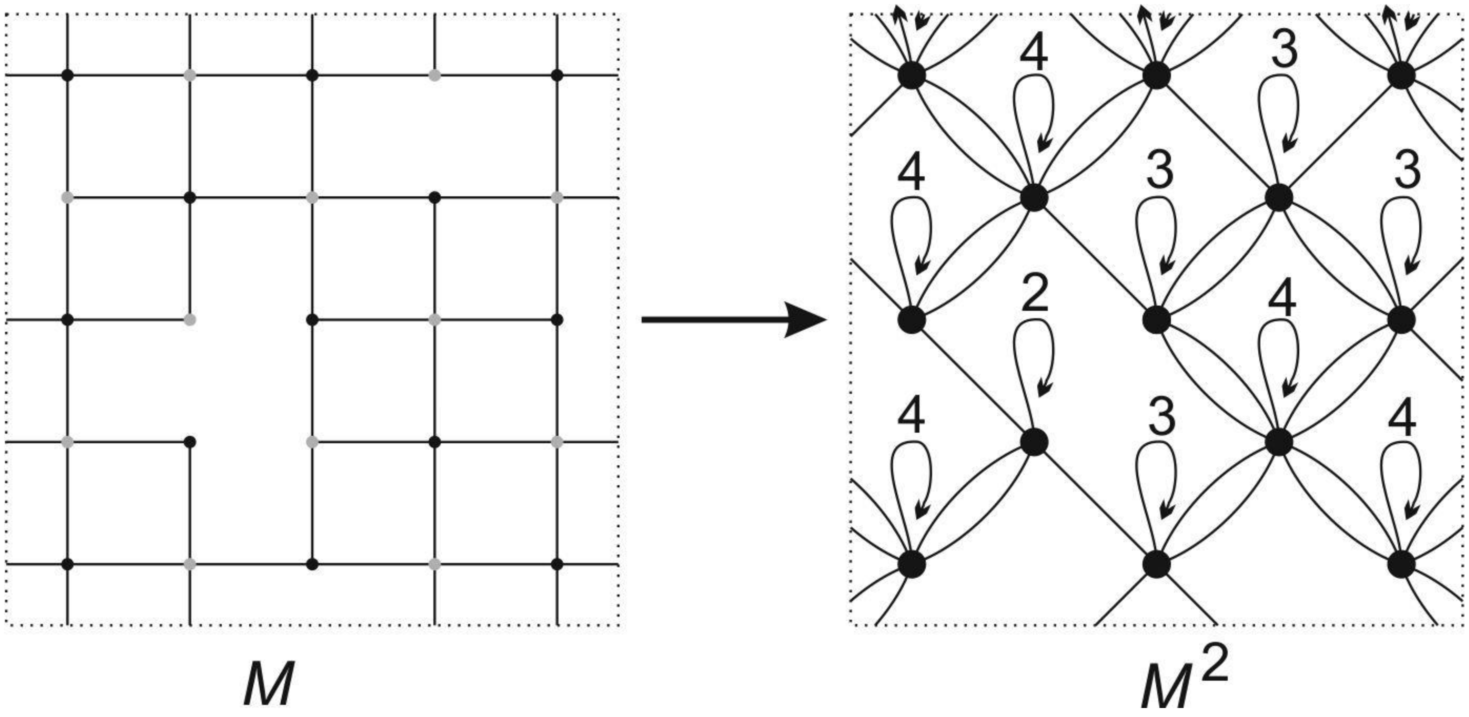
$$\boxed{(S^k)_{ab} = \frac{(M^k)_{ab}}{\lambda^k}\frac{\psi_b}{\psi_a}}$$

Renormalization (being **scale-free:** discretization independent**)**

We can change not only time scale, but also spatial

$$\left(\left(S^{\mathrm{MERW}(M)}\right)^{l}\right)_{ij} = \sum_{\gamma_1,\ldots,\gamma_{k-1}} \frac{M_{i\gamma_1}}{\lambda}\frac{\psi_{\gamma_1}}{\psi_i}\cdot\frac{M_{\gamma_1\gamma_2}}{\lambda}\frac{\psi_{\gamma_2}}{\psi_{\gamma_1}}\cdot\ldots\cdot\frac{M_{\gamma_{k-1}\gamma_k}}{\lambda}\frac{\psi_j}{\psi_{\gamma_{k-1}}} = \frac{(M^l)_{ij}}{\lambda^k}\frac{\psi_{\gamma_k}}{\psi_{\gamma_0}} = \left(S^{\mathrm{MERW}(M^l)}\right)_{ij}$$

Usually not true for GRW

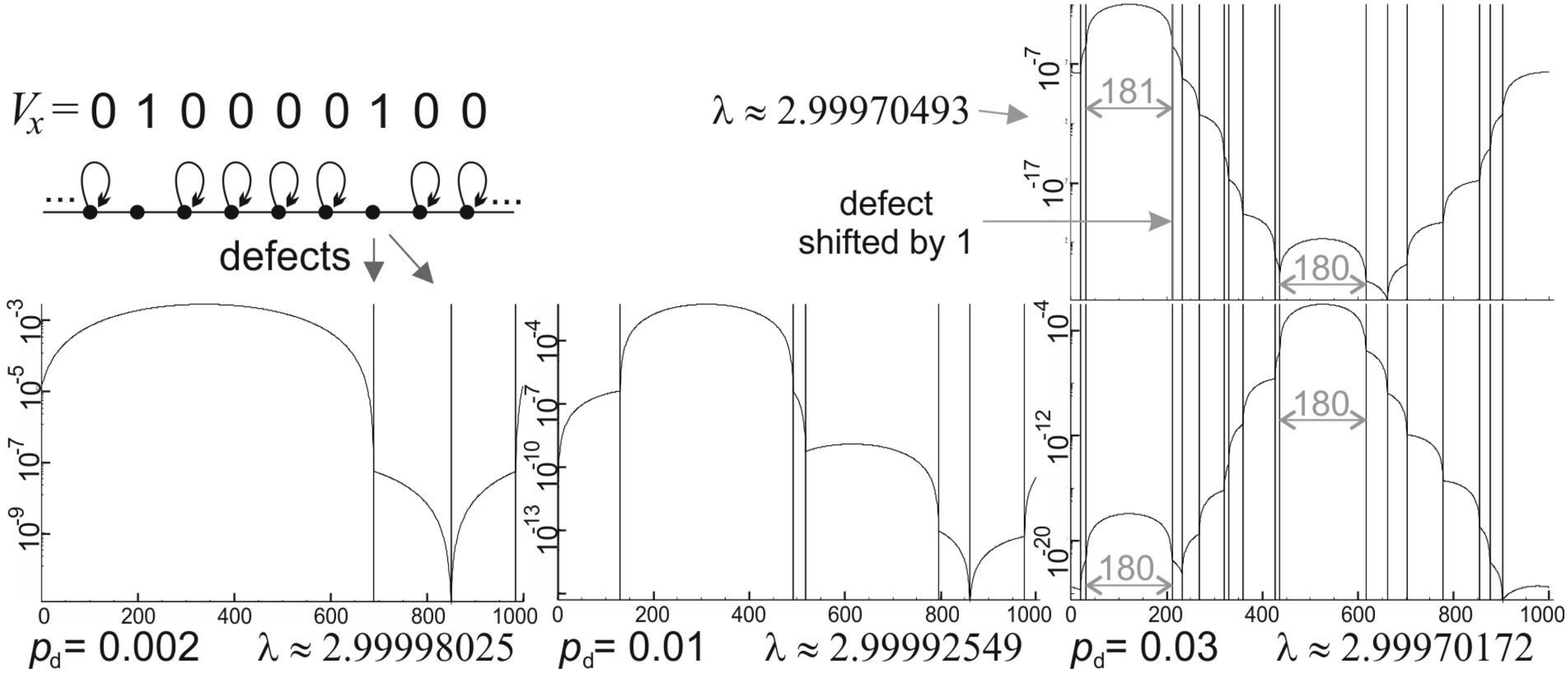


GRW: stationary probability $\boxed{\propto d_i = \sum_j M_{ij}}$

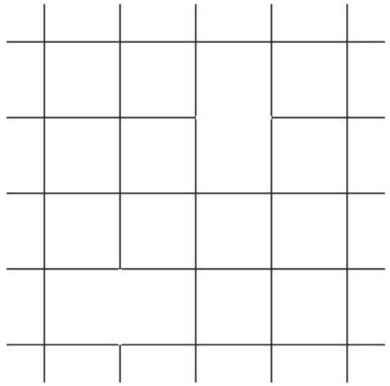
MERW: stationary probability $\boxed{\propto \psi^2}$ where $\boxed{M\psi = \lambda\psi}$ for largest λ

Defected	$(\lambda\psi)_x = (M\psi)_x = \psi_{x-1} + (1 - V_x)\psi_x + \psi_{x+1}$	$/-3\psi_x \quad /\cdot -1$
1D lattice	$E\psi_x = -(\psi_{x-1} - 2\psi_x + \psi_{x+1}) + V_x\psi_x$	for smallest $E = 3 - \lambda$

Nonlocal – depends on the whole graph!



Idealized situation: **defected lattice** (cyclic boundary conditions) →

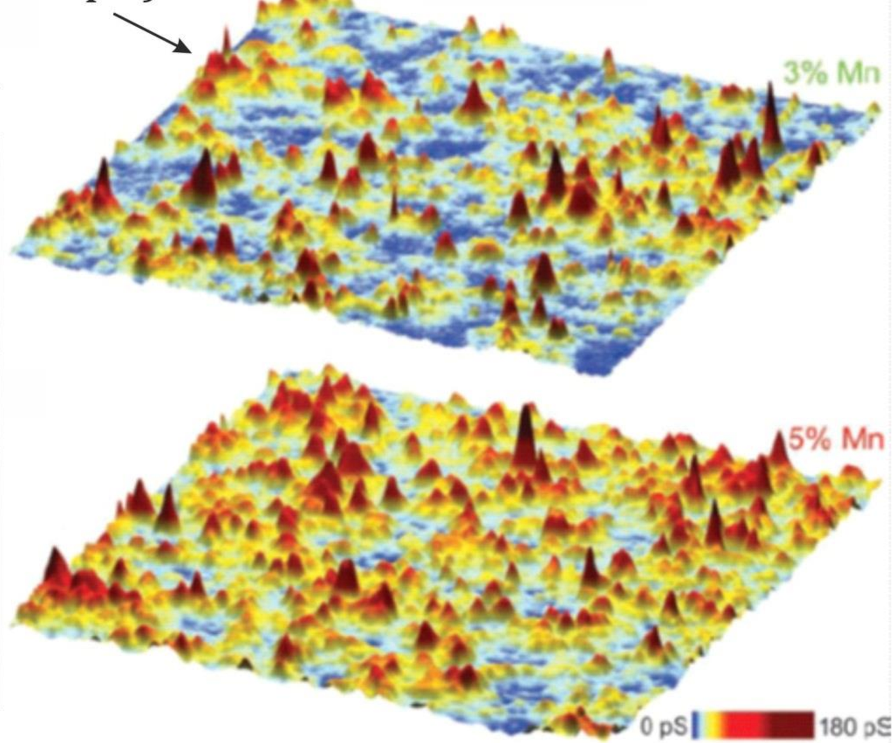


“Natural” stochastic <u>choice</u> (“drunken sailor”):
Each outgoing edge is equally probable (GenericRW)

Bose-Hubbard Hamiltonian (→ **Schrödinger**) <u>for single particle</u>:

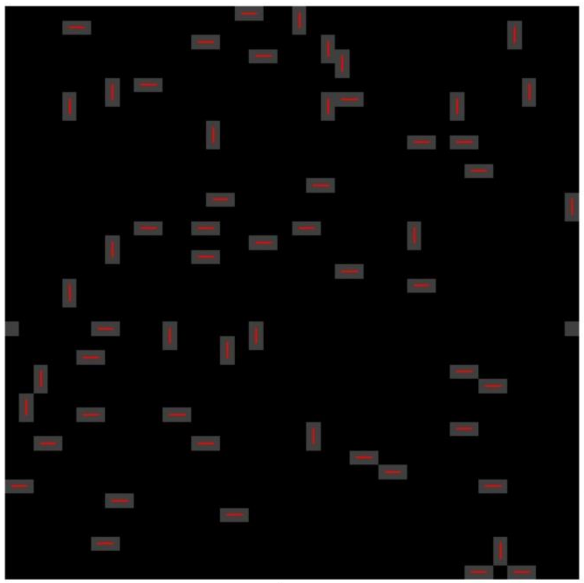
$$\hat{H} = -t \sum_{(i,j)\in\mathcal{E}} \left(\hat{a}_j^{+}\hat{a}_i + \hat{a}_i^{+}\hat{a}_j\right) = -t \cdot \text{"adjacency matrix"}$$

STM measurenments of electron density for $Ga_{1-x}Mn_x As$ (20pA) 70 x 70nm

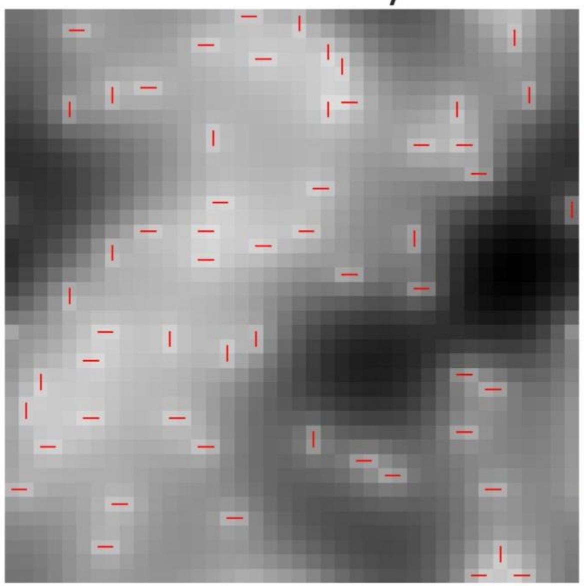


GRW

ground state density
of Bose-Hubbard / MERW



conductor



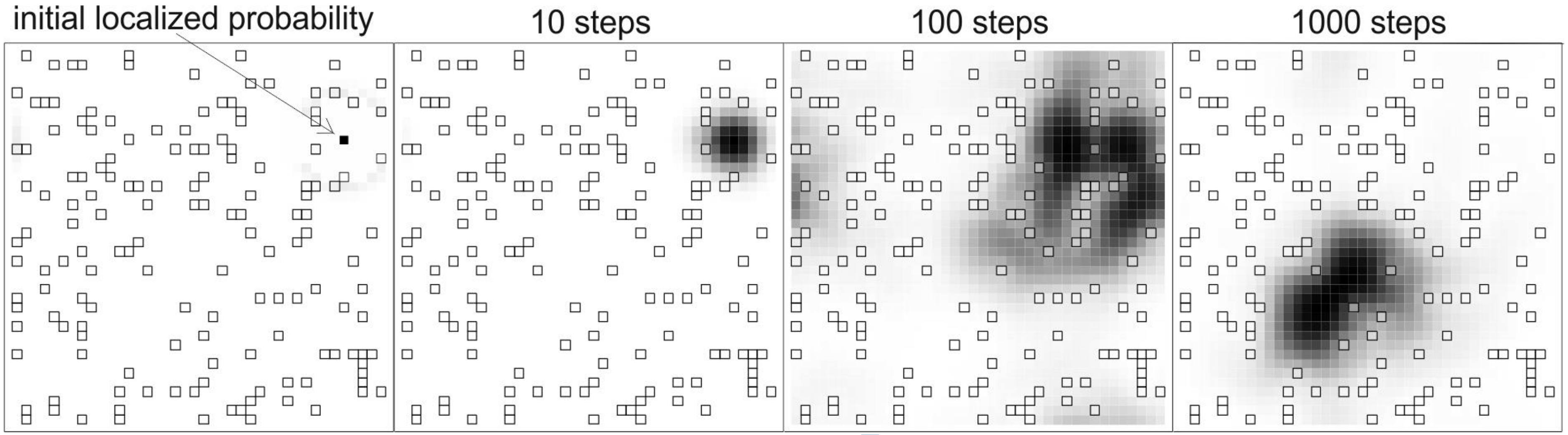
insulator

Discrepancy source: **GRW only approximates maximal uncertainty principle**

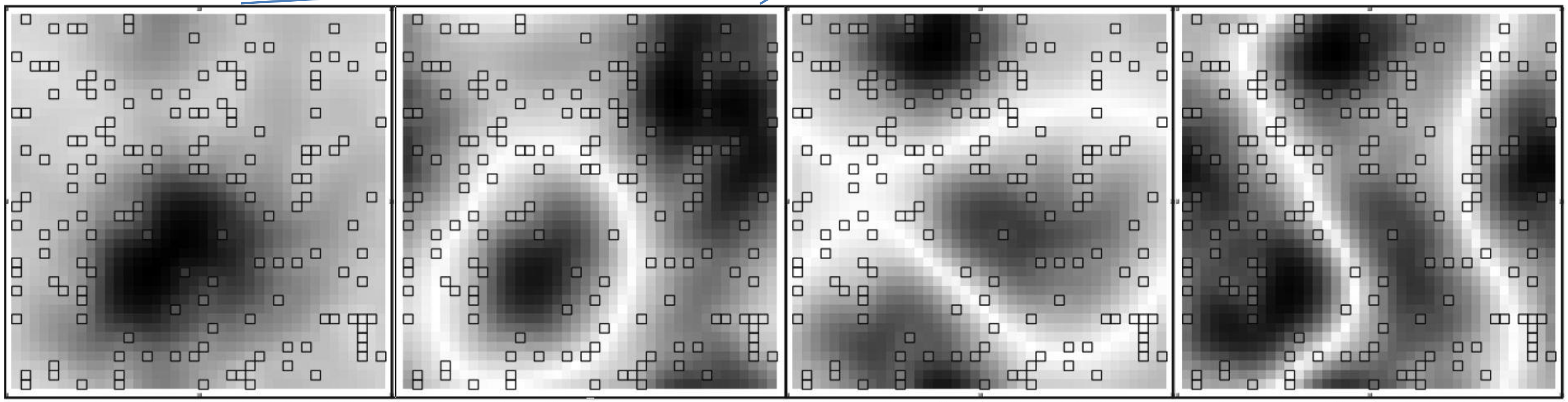
MERW evolution:

$$\left(S^{\mathrm{M}}\right)^t_{ij} = \frac{(M)^t_{ij}}{\lambda_0^t}\frac{\psi_{0,j}}{\psi_{0,i}} = \left(\sum_k \left(\frac{\lambda_k}{\lambda_0}\right)^t \varphi_{k,j}\psi_{k,i}\right)\frac{\psi_{0,j}}{\psi_{0,i}}$$

First "stochastic shift" toward **near** (overlapping) eigenvectors (sub-diffusion), then "deexcitate" toward nearer **ground state** (super-diffusion)



Eigenvectors $|\psi_k|$:



$k = 0 \quad \lambda \approx 4.942458$ $k = 1 \quad \lambda \approx 4.922788$ $k = 2 \quad \lambda \approx 4.918159$ $k = 3 \quad \lambda \approx 4.906692$

Add potential to emphasize some scenarios: **Boltzmann distribution**

maximizes entropy while fixed sum of energies (minimizes free energy)

$$\max_{(p_i):\sum_i p_i=1} \left(\sum_i p_i \ln(1/p_i) - \sum_i p_i E_i\right) = \ln\left(\sum_i e^{-E_i}\right) \qquad \text{for} \qquad p_i \propto e^{-E_i}$$

Original MERW: $A_{ij} \in \{0,1\}$

Maximize entropy - uniform probability distribution among paths

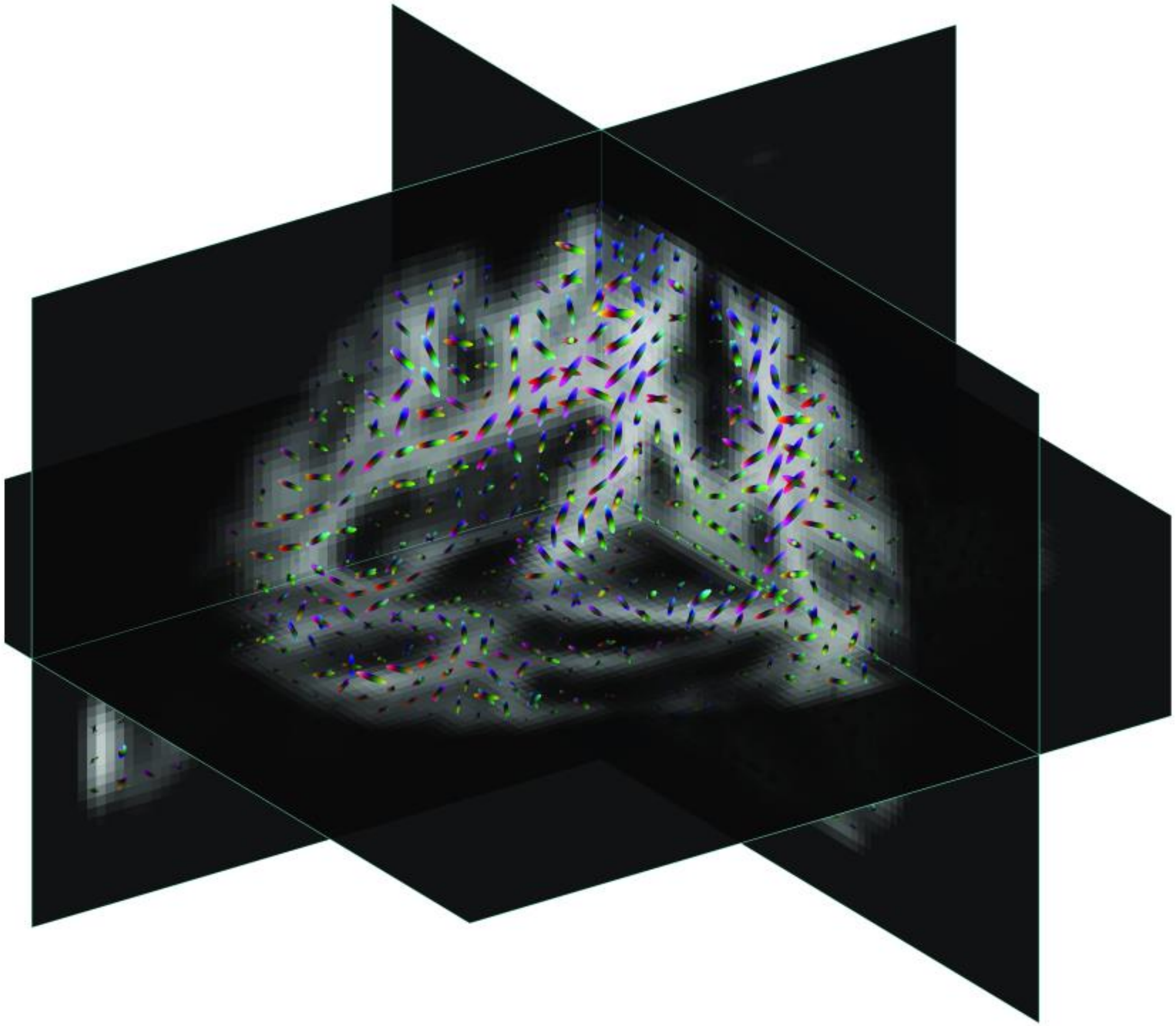
Generally: minimize free energy – Boltzmann distribution among paths:

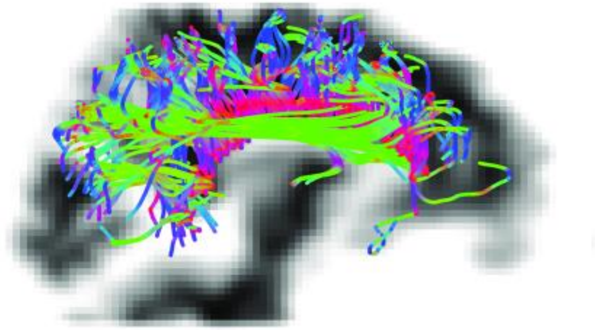
$M_{ij} = A_{ij} e^{-\beta V_{ij}} \in [0,\infty)$ Energy of path γ: $V_{\gamma_0\gamma_1} + V_{\gamma_1\gamma_2} + .. + V_{\gamma_{l-1}\gamma_l}$

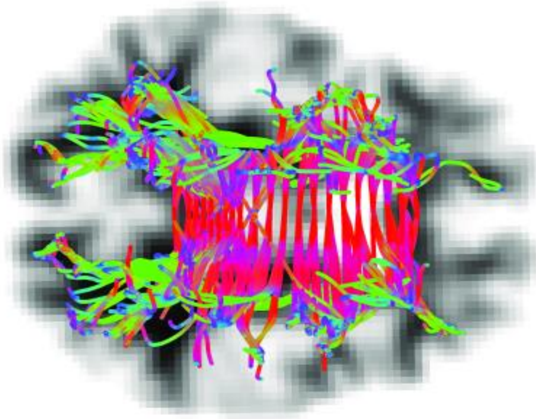
$$S_{\gamma_0\gamma_1} S_{\gamma_1\gamma_2} \ldots S_{\gamma_{l-1}\gamma_l} = \frac{M_{\gamma_0\gamma_1} \ldots M_{\gamma_{l-1}\gamma_l}}{\lambda^l} \frac{\psi_{\gamma_l}}{\psi_{\gamma_0}} = \frac{e^{-\beta\left(V_{\gamma_0\gamma_1} + V_{\gamma_1\gamma_2} + .. + V_{\gamma_{l-1}\gamma_l}\right)}}{\lambda^l} \frac{\psi_{\gamma_l}}{\psi_{\gamma_0}}$$

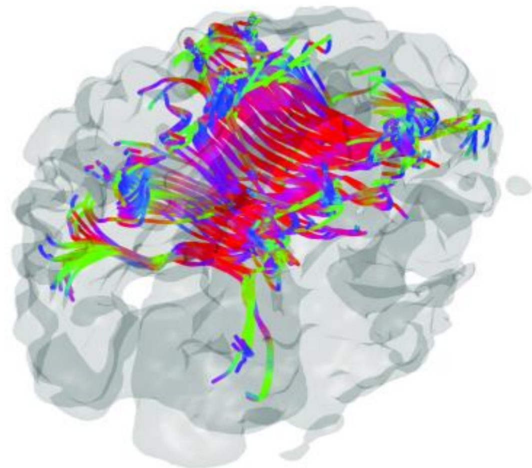
Alternative view: $\boldsymbol{M_{ij}}$ **is the number of edges** (not necessarily 1, integer)

Simultaneous Multi-Scale Diffusion Estimation an
Tractography Guided by Entropy Spectrum Pathways
Vitaly L. Galinsky and Lawrence R. Frank, IEEE
Transactions on Medical Imaging (2014)

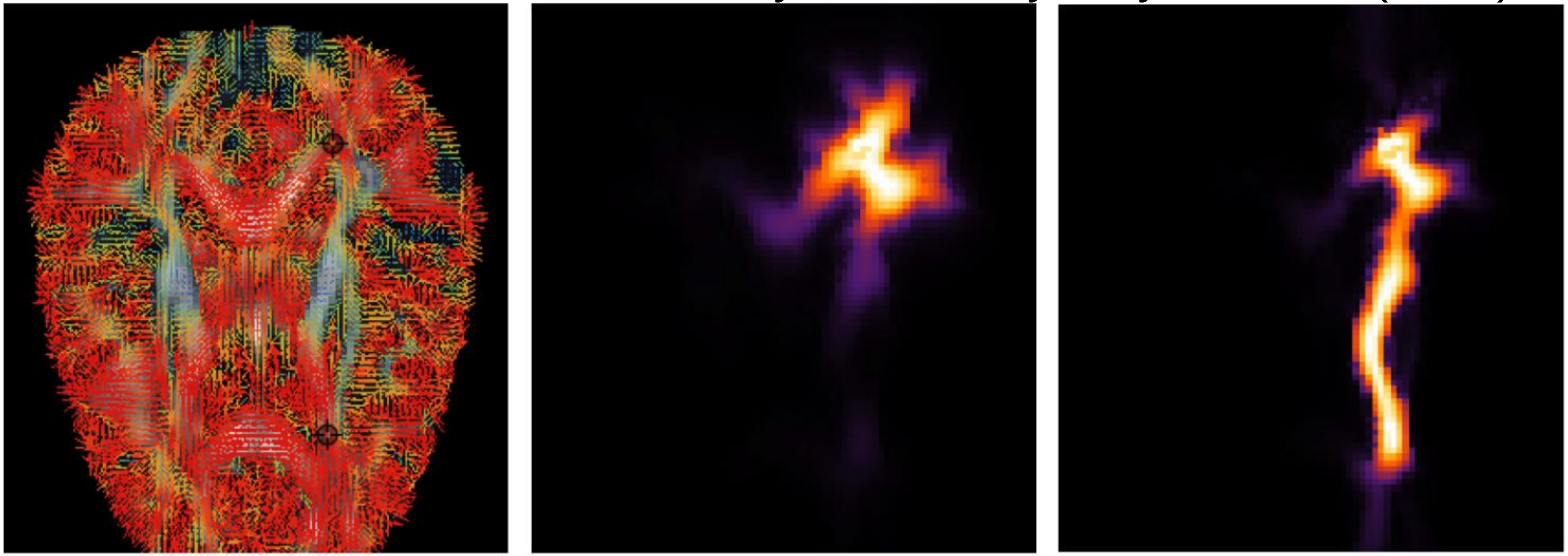






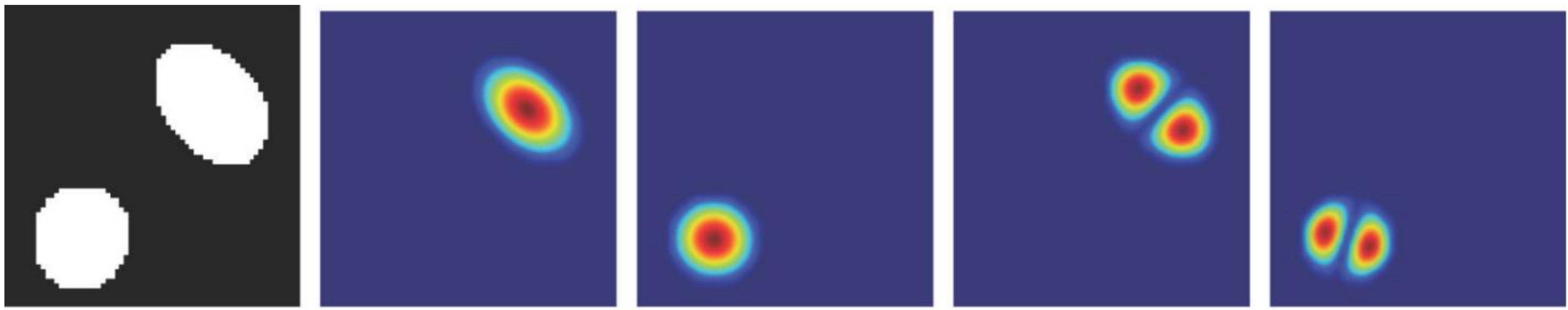


Information pathways in a disordered lattice,
Lawrence R. Frank 1,2,* and Vitaly L. Galinsky, Phys. Rev. E (2014)



(a) DT-MRI data (b) Fiber tracking with GRW (c) Fiber tracking with ESP

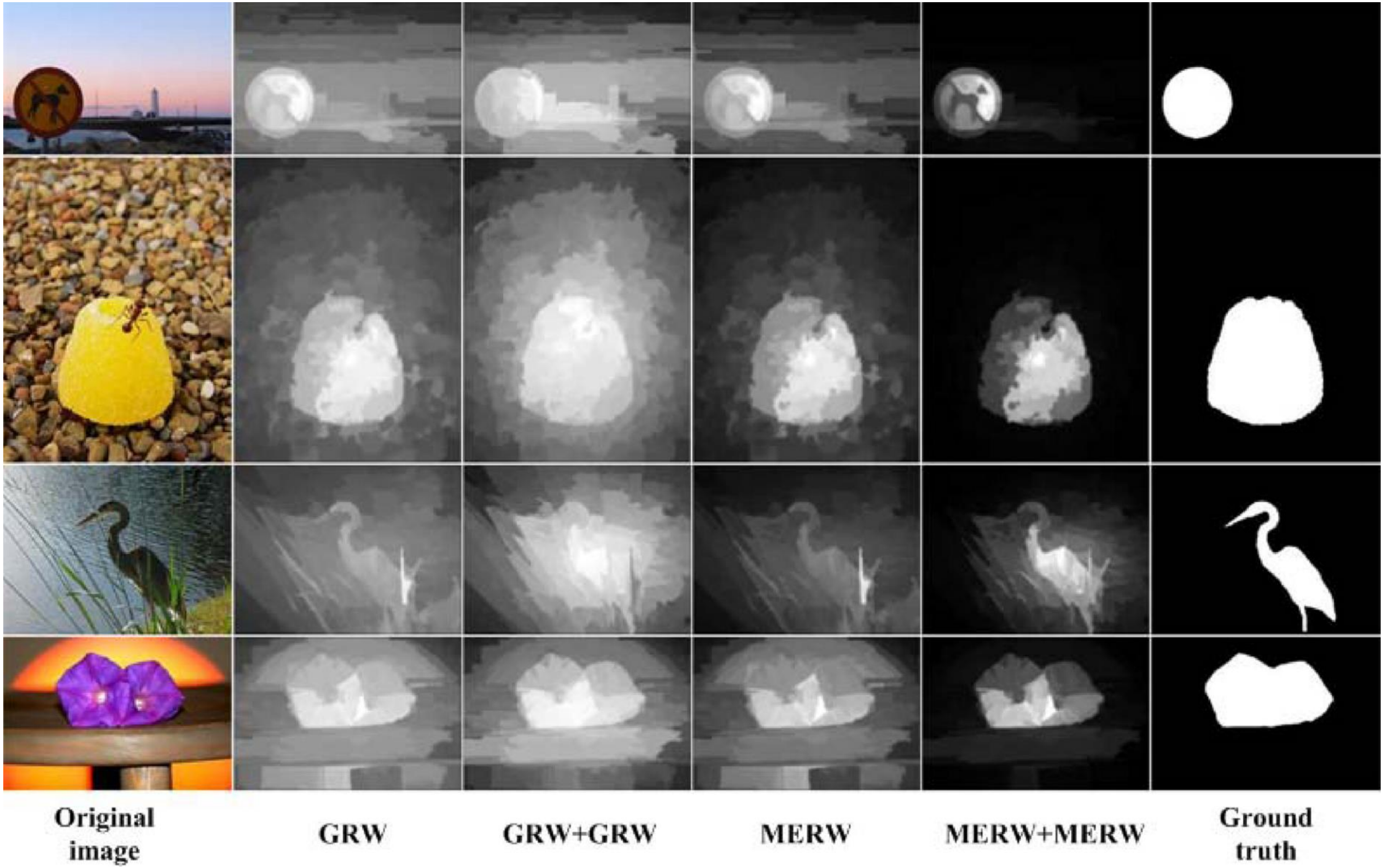
Entropy Spectrum Pathways (ESP): generalization to multiple dominant eigenvectors (entropy wells)



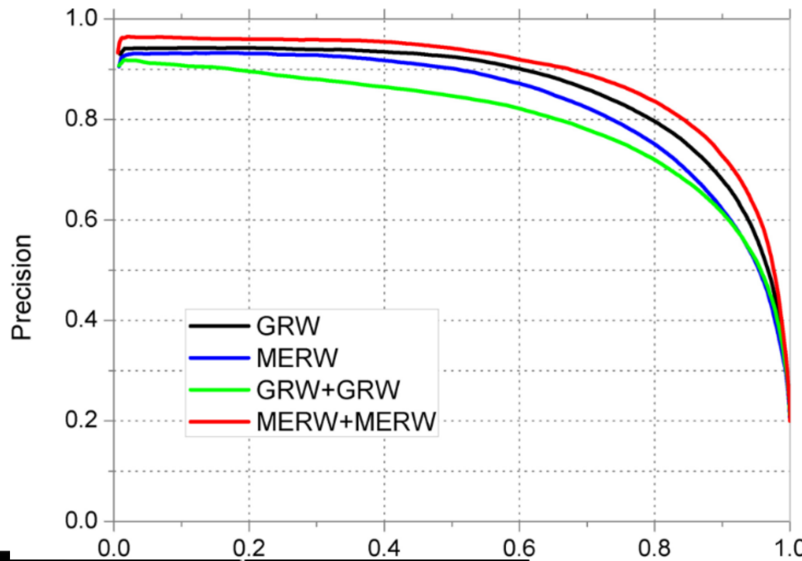
(a) Q_{ij} (b) e_1 (c) e_2 (d) e_3 (e) e_4

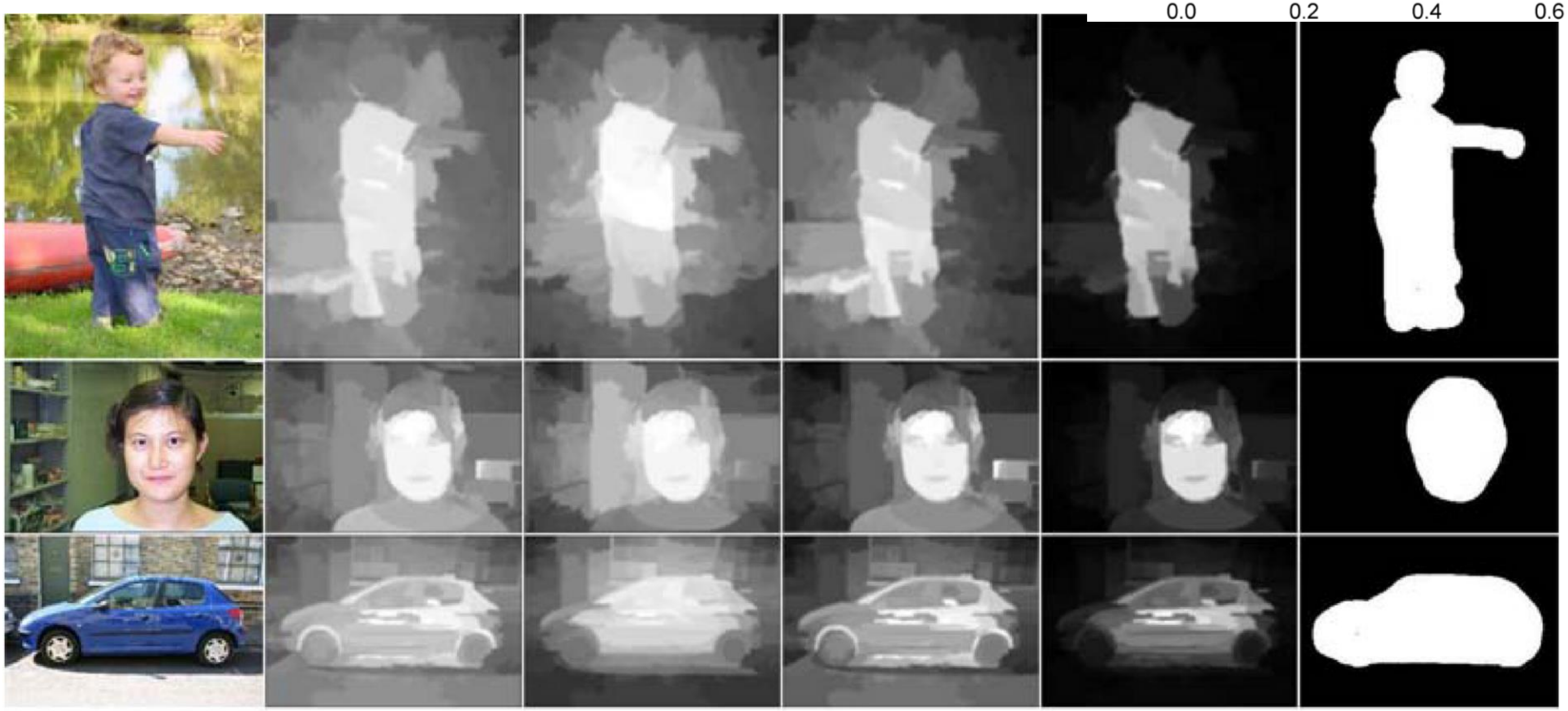
Using MERW properties (localization) for various applications

JG Yu, J Zhao, J Tian, Y Tan, *Maximal Entropy Random Walk for Region-Based Visual Saliency* (IEEE, 2014)



- divide picture into regions (8x8 blocks, "superpixels")
- create graph among regions using similarities as weights ($w_{ij} = \exp(-d(r_i, r_j))$),
- saliency map is the stationary probability distribution of GRW or MERW





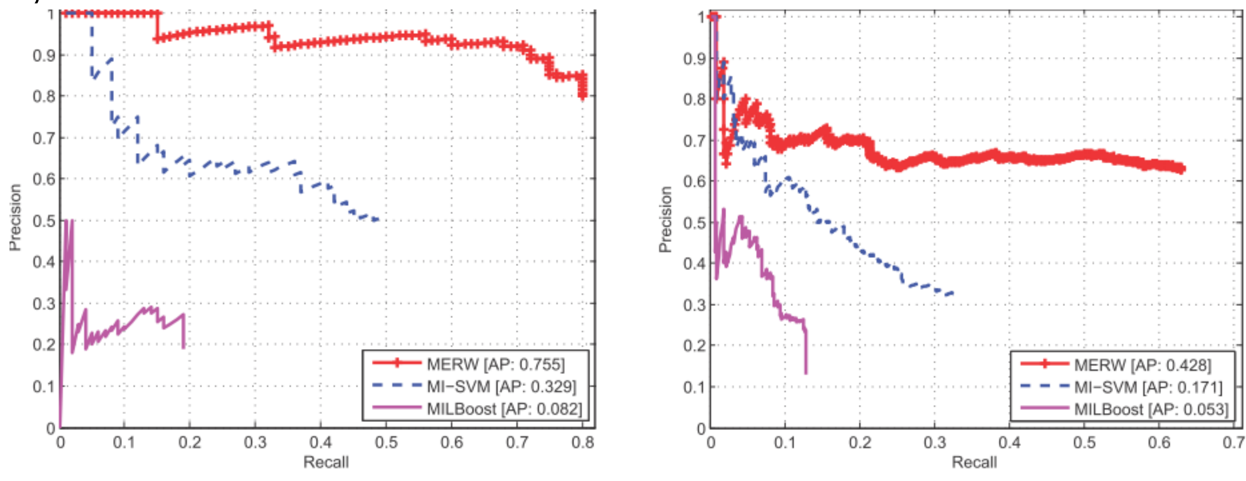
Original image | GRW | GRW+GRW | MERW | MERW+MERW | Ground truth

WEAKLY SUPERVISED OBJECT LOCALIZATION VIA MAXIMAL ENTROPY RANDOM WALK,

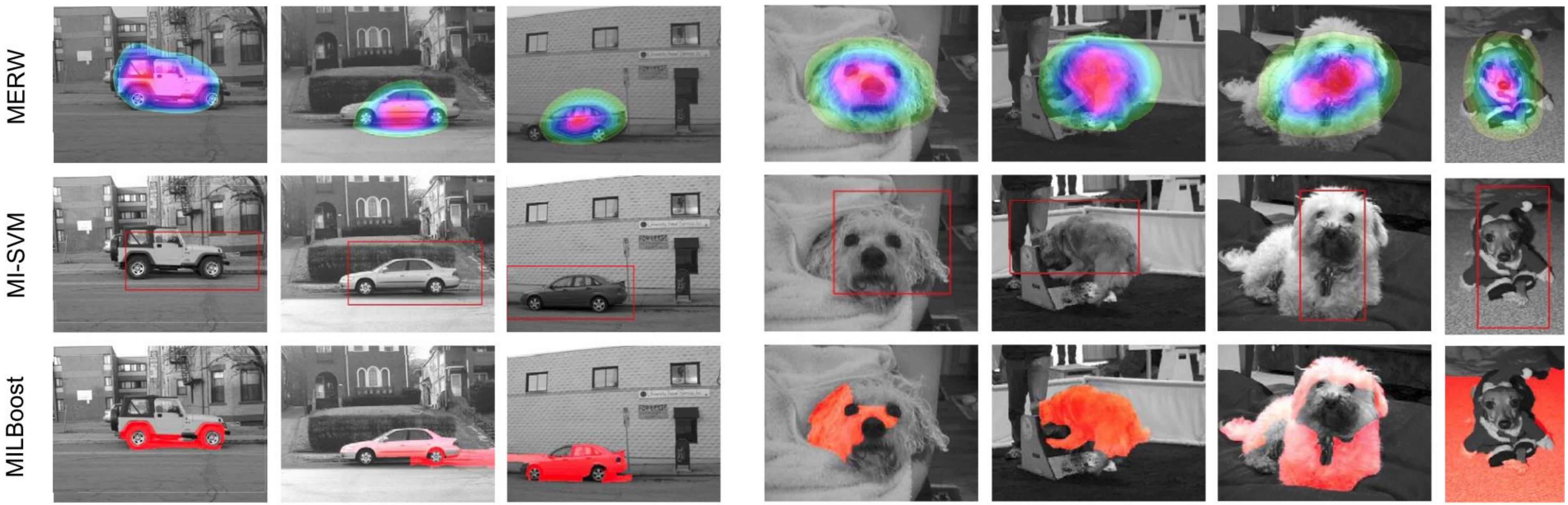
Liantao Wang, Ji Zhao, Xuelei Hu, Jianfeng Lu, IEEE ICIP 2014

Divide the picture into regions and use SVM to evaluate weights of features (w_i) for different objects (e.g. car, dog)

$$a_{ij} = \begin{cases} e^{\gamma(w_i+w_j)}, & \text{if} \quad z_j \in \mathcal{N}_k(z_i) \\ 0, & \text{otherwise} \end{cases}$$



(a) PittCar (b) ‘dog’



Centrality (graph theory, http://en.wikipedia.org/wiki/Centrality): **indicators which identify the most important vertices within a graph.**

Examples (for the same graph):

A) Degree centrality
($e.g.\ C(v) \propto \deg(v)$ $-$ GRW),

B) Closeness centrality
($e.g.\ C(v) \propto \sum_{w \neq v} 1/d(v,w)$),

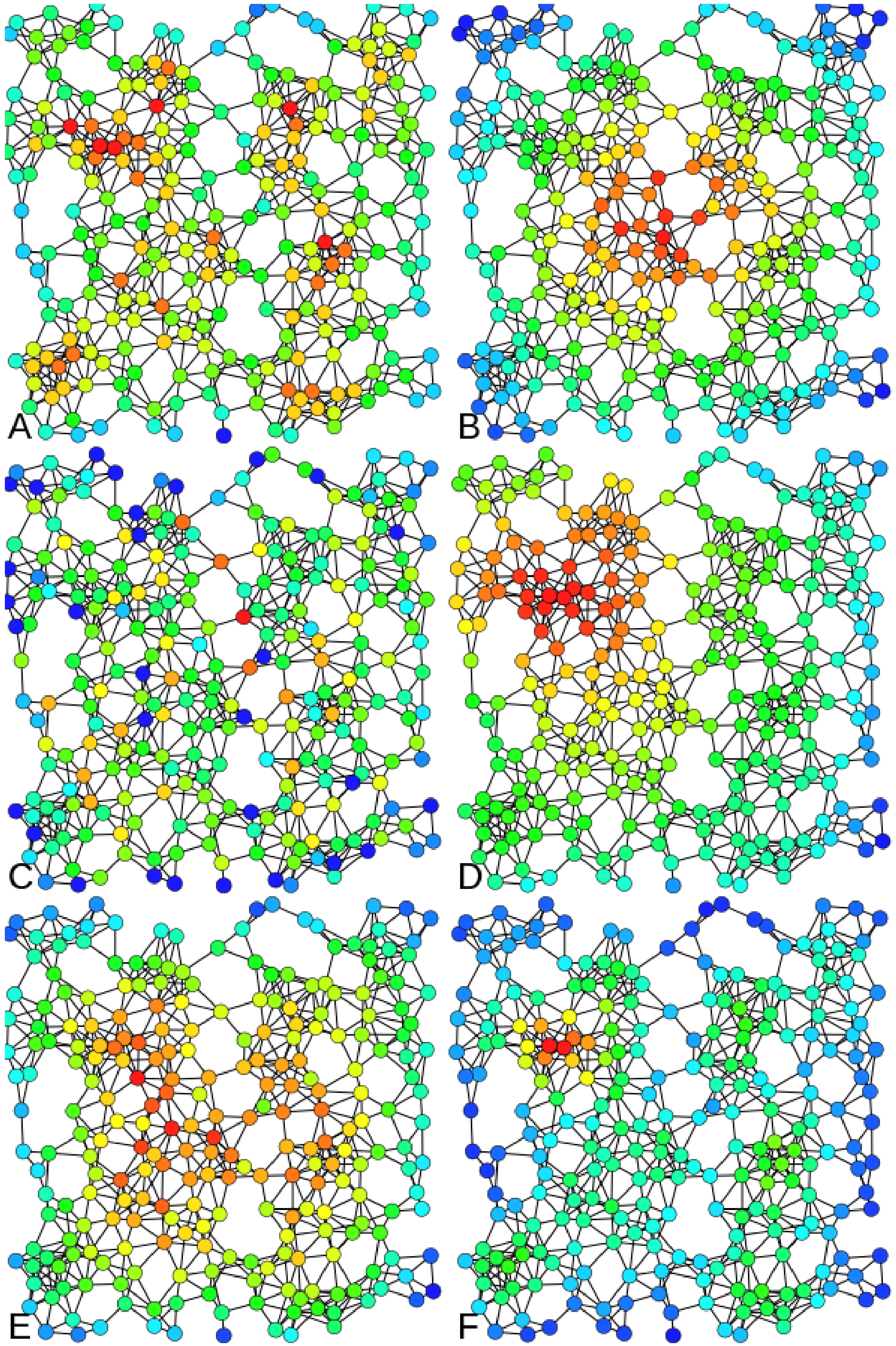
C) Betweenness centrality
(how many shortest paths go through v)

D) **Eigenvector centrality** **(MERW-like)**,

E) Katz centrality (e.g. PageRank),

F) Alpha centrality.

Drawing 2D diagrams for graphs:
positions from two high eigenvectors
(of M or Laplacian: $L = \mathrm{diag}(\deg(i)) - M$)



Delvenne, J.-C. & Libert, A.-S. *Centrality measures and thermodynamic formalism for complex networks,* Phys. Rev. E 83, 046117 (2011).

(e.g. Google) PageRank (GRW) → Entropy Rank (MERW)

(α = Pr(going to a random page), $E = e^{-U_0}$ weight out of the graph edges)

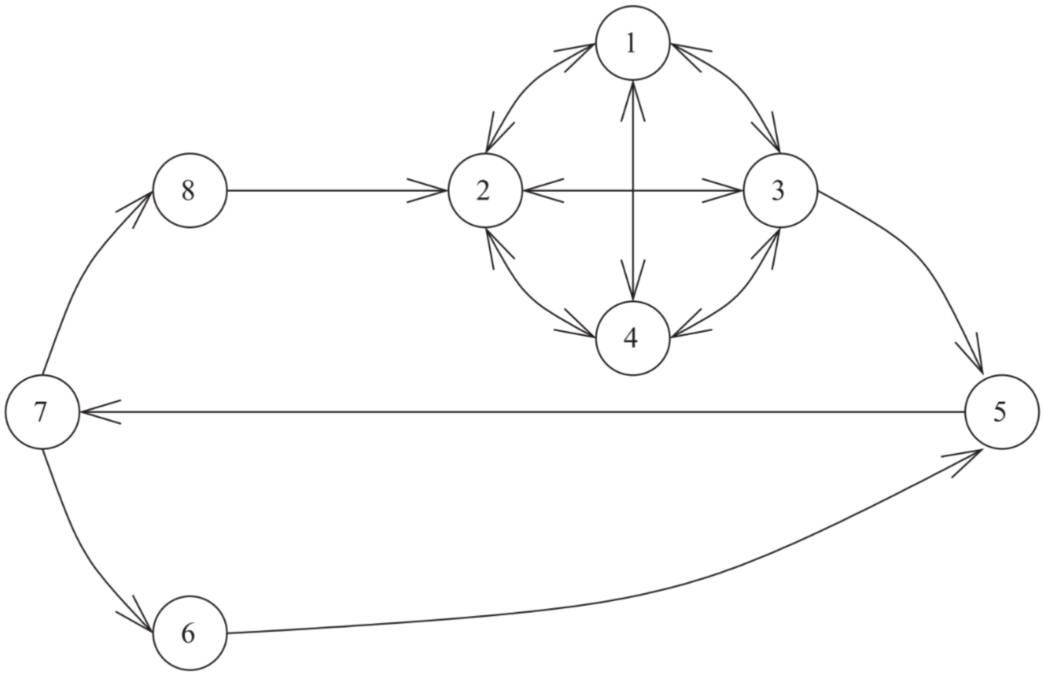
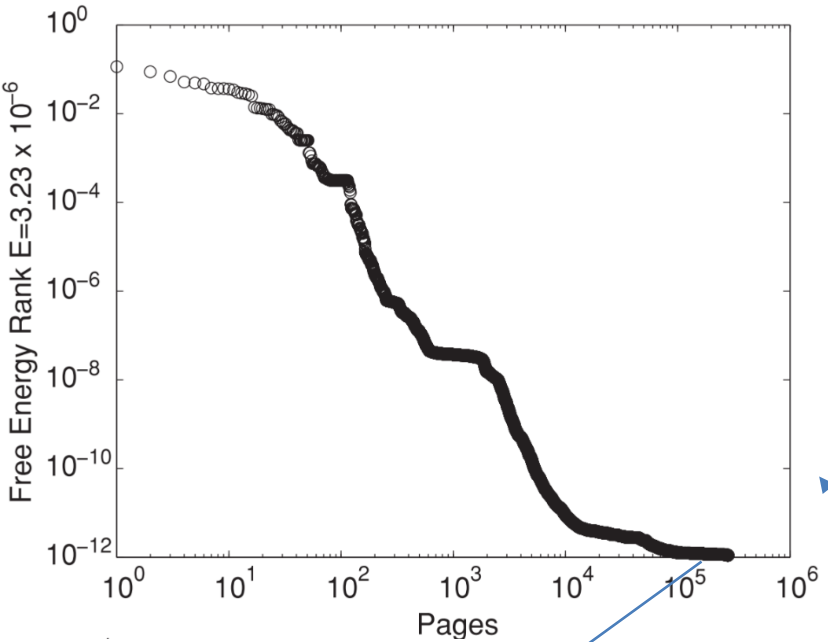


TABLE I. PageRank, free-energy rank, and entropy rank for the network of Fig. 1.

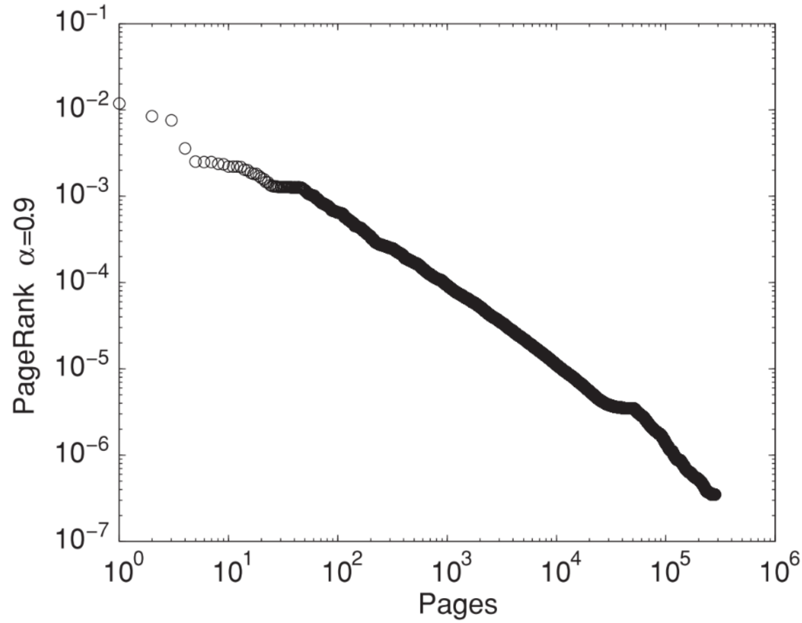
Vertex	PageRank ($\alpha = 1$)	PageRank ($\alpha = 0.9$)	Entropy rank	Free-energy rank ($E = 0.03$)
1	0.1705	0.1549	0.2464	0.2400
2	0.2045	0.1965	0.2487	0.2458
3	0.1818	0.1644	0.2487	0.2460
4	0.1705	0.1549	0.2464	0.2400
5	0.0909	0.1035	0.0032	0.0099
6	0.0455	0.0601	0.0001	0.0019
7	0.0909	0.1057	0.0032	0.0076
8	0.0455	0.0601	0.0031	0.0087

- vertex 8 becomes more interesting than 6 (pointing to "good pages"),
- cliques are swelling (localization) – problem with "link farms" …

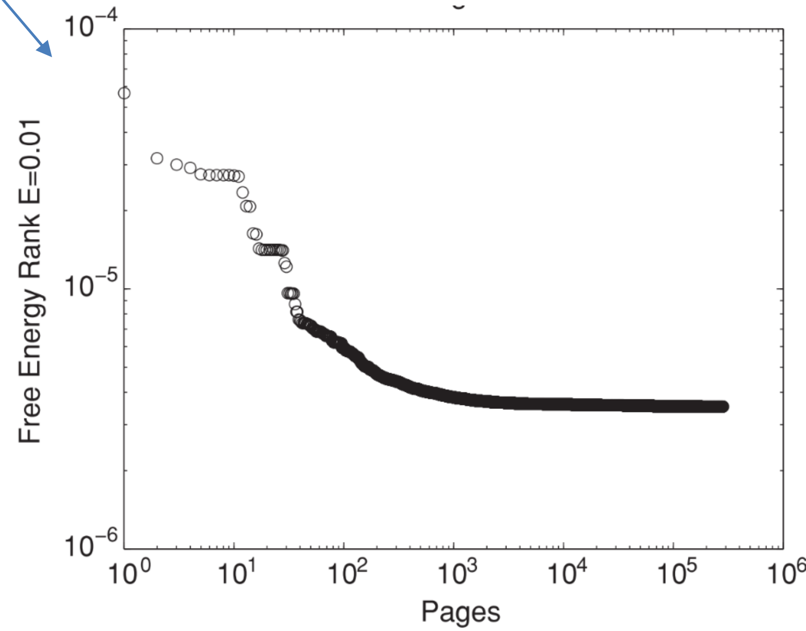
Experiments on "289 000 – node piece of the Stanford web (http://www.kamvar.org/)"



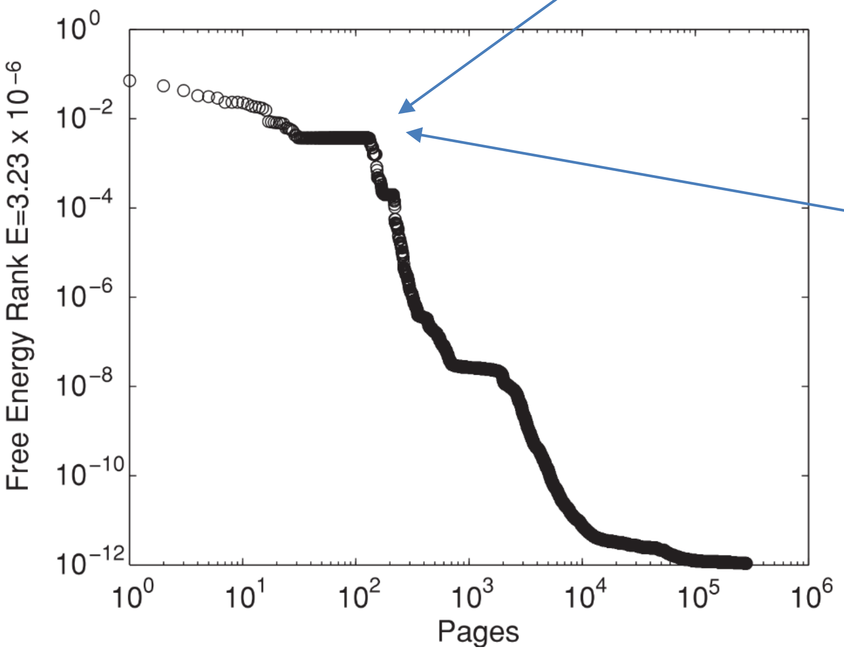
PageRank



High E FER
(good for finding
best pages)



low H FER



low H FER
vertex with added
100 vert. clique
("farm link")
$200\ 000th \rightarrow 627th$
(plateau → clique ?)

Mean first-passage time (MFPT) (e.g. for community detection)

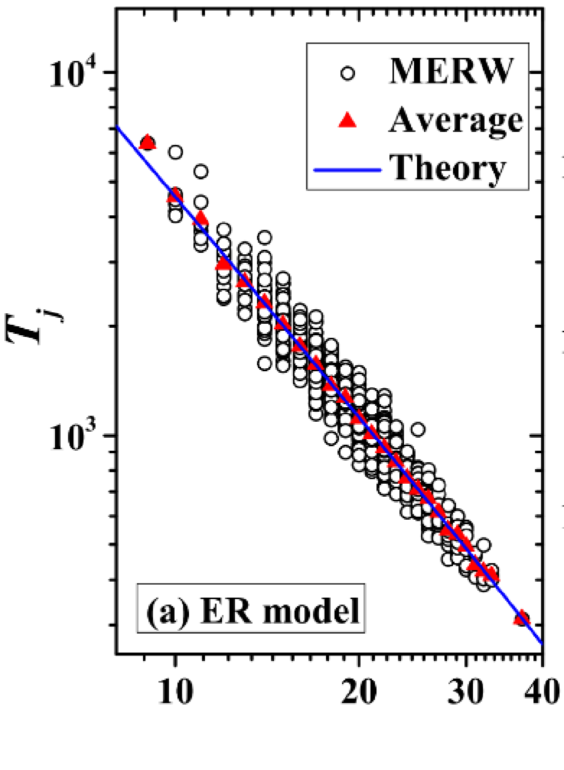
M_{ij} – expected minimal time to reach vertex j starting from i.

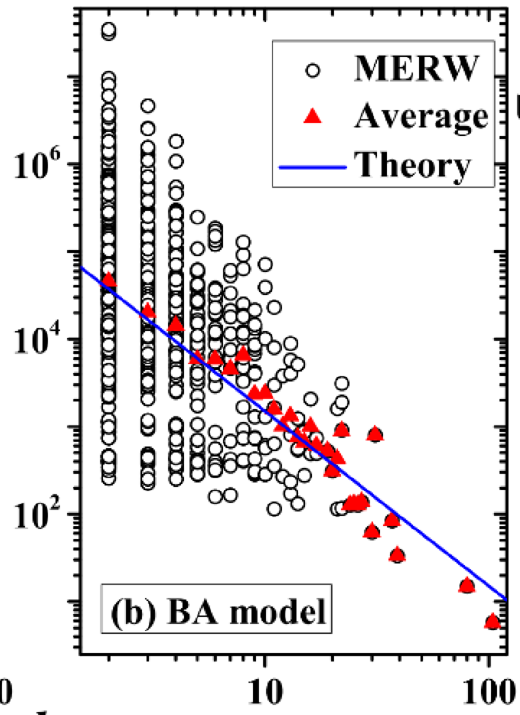
Y. Lin, Z. Zhang, *Mean first-passage time for maximal-entropy random walks in complex networks* (Nature, 2014)

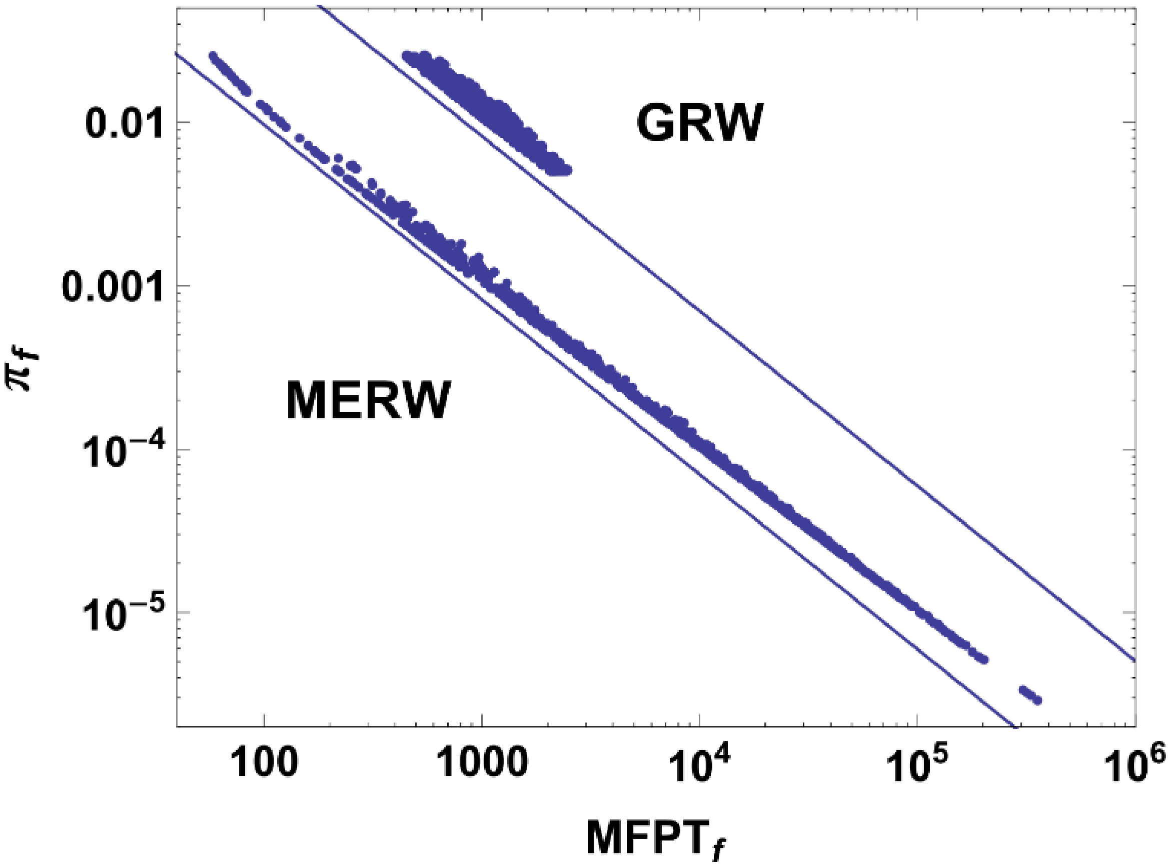
Erdős–Rényi (ER): $\Pr(\to v_j) = \text{const}$

Barabási–Albert (BA): $\Pr(\to v_j) \propto k_j$

(scale-free : $P(k) \sim k^{-\gamma}$)







1000 vertices

J. Ochab, *Maximal-entropy random walk unifies centrality measures* (Phys. Rev. E, 2012)

SimRank: measure how similar two vertices are

G. Jeh and J. Widom. *Simrank: a measure of structural-context similarity* (KDD 2002)

$s(a,b) = \frac{C}{|N(a)||N(b)|}\sum_{x\in N(a)}\sum_{y\in N(b)} s(x,y)$ $\quad$ $(1\ if\ a=b,\ 0\ if\ I(a)\cap I(b)=\emptyset)$

can be expressed by Expected$-f$ Meeting Distance (EMD) of two walkers (a,b)

$s'(a,b) = \sum_{t:(a,b)\rightsquigarrow(x,x)} P[t]\, f\big(l(t)\big)$ $\quad$ for $f(z)=z$ $\quad$ or $f(z)=C^z$

$P[t]$ - GRW probability of path t

Link prediction – which new interactions (links) are likely to occur?

Predicting evolution, suggesting connections, finding weak/fake links

The more similar they are, the more likely they will connect

Li, R. H., Yu, J. X. & Liu, J. *Link prediction: the power of maximal entropy random walk* (ACM conference, 2011):

Replace GRW with MERW in $P[t]$, getting $S(a,b) = \frac{C\psi_a\psi_b}{\lambda^2}\sum_{x\in N(a)}\sum_{y\in N(b)}\frac{S(x,y)}{\psi_x\psi_y}$

MERW – more distinctive, scale-free (does not depend on discretization)

27 link prediction methods (AUC: the higher the better), “ME” – maximal entropy

SM	ER	BA	SW	USAir	C.ele	Yeast	Power	NetSci	GrQc	HepPh	HepTh
CTT	0.710	0.750	0.791	0.847	0.784	0.709	0.713	0.917	0.520	0.523	0.525
CTME	0.720	0.746	0.745	0.855	0.798	0.501	0.501	0.866	0.556	0.645	0.534
CK	0.805	0.883	0.804	0.856	0.809	0.715	0.501	0.799	0.513	0.501	0.513
MECK	0.940	0.981	0.845	0.936	0.856	0.757	0.501	0.975	0.517	0.501	0.503
NCK	0.502	0.501	0.501	0.708	0.706	0.501	0.501	0.501	0.503	0.508	0.501
NMECK	0.903	0.983	0.982	0.931	0.969	0.710	0.501	0.971	0.623	0.750	0.675
DK	0.835	0.813	0.983	0.836	0.838	0.829	0.764	0.965	0.501	0.605	0.593
MEDK	0.999	0.983	0.998	0.991	0.971	0.749	0.812	0.963	0.739	0.735	0.746
NDK	0.786	0.711	0.956	0.920	0.778	0.731	0.857	0.908	0.531	0.530	0.530
NMEDK	**0.999**	**0.983**	**0.998**	**0.997**	**0.978**	**0.970**	0.857	**0.996**	**0.739**	0.755	**0.758**
RK	0.851	0.907	0.973	0.898	0.887	0.803	0.864	0.624	0.632	0.608	0.561
MERK	0.999	0.983	0.998	0.981	0.949	0.812	0.812	0.963	0.618	0.745	0.735
NRK	0.504	0.501	0.501	0.719	0.501	0.703	0.806	0.501	0.501	0.508	0.504
NMERK	0.999	0.983	0.998	0.983	0.975	0.968	0.857	0.986	0.739	0.755	0.756
MENK	0.999	0.983	0.998	0.936	0.975	0.799	0.812	0.963	0.618	0.730	0.746
NNK	0.503	0.501	0.501	0.819	0.501	0.705	0.806	0.501	0.501	0.508	0.504
NMENK	0.999	0.983	0.998	0.983	0.965	0.965	0.857	0.996	0.739	0.755	0.752
PD	0.926	0.974	0.953	0.971	0.866	0.887	0.857	0.722	0.666	0.618	0.628
MEPD	0.999	0.976	0.998	0.993	0.964	0.968	0.857	0.913	0.739	0.755	0.758
PDM	0.805	0.764	0.957	0.972	0.798	0.886	0.857	0.874	0.616	0.660	0.530
MEPDM	0.999	0.983	0.998	0.990	0.976	0.970	0.857	0.996	0.739	0.755	0.758
SR	–	–	–	0.905	0.860	–	–	0.955	–	–	–
MESR	–	–	–	0.960	0.876	–	–	0.963	–	–	–
CN	0.884	0.782	0.501	0.386	0.971	0.752	0.802	0.961	0.617	0.623	0.635
AA	0.886	0.781	0.501	0.409	0.975	0.793	0.806	0.969	0.623	0.630	0.638
HPLP+	0.983	0.971	0.978	0.979	0.974	0.965	**0.886**	0.984	0.725	0.753	0.732
SRW	0.991	0.977	0.989	0.983	0.972	0.967	0.863	0.983	0.731	**0.760**	0.754

Kernel between G and G': $k(G,G') = q_\times^T \cdot \left(\sum_{k\geq 0} \mu(k)\, W_\times^k\right) \cdot p_\times$ e.g. $(1-\lambda W_\times)^{-1}$ or $e^{\lambda W_\times}$

NMEDK – normalized maximal entropy heat diffusion kernel, **NMERK** - ...Laplacian kernel

$S_{ij}^{GRW} = \frac{M_{ij}}{\deg(i)}$ $\pi_i^{GRW} \propto \deg(i)$ $(S^{MERW})_{ij}^t = \frac{(M)^t{}_{ij}}{\lambda^k}\frac{\psi_j}{\psi_j}$ $\pi_i^{MERW} \propto \psi_i^2$

GRW Laplacian $(M_{ii}=0)$: $\Delta_{ij} = -\mathrm{L}_{ij} = M_{ij} - \deg(i)\cdot\delta_{ij}$ $\left(w^T \mathrm{L} w = \sum_{\{i,j\}\in E}\left(w_i - w_j\right)^2\right)$

In analogy to discretized continuous Laplacian: $(\partial_{xx} w)(x) \approx w(x-1) - 2w(x) + w(x+1)$

Or relaxation of capacitor network: $\frac{d}{dt} Z_i(t) = \sum_{j:\, i\sim j}\left(Z_j(t) - Z_i(t)\right)$.

General Laplacian ("**continuity equation**": $\forall_i \sum_j L_{ij} = 0$, $M_{ij} = M_{ji} \Rightarrow \Pr(i,j) = \Pr(j,i)$)**:**

(const $\cdot$) $\Delta_{ij} = \left(\Pi(S - \mathbf{1})\right)_{ij} = \Pr(i,j) - \Pr(i)\cdot\delta_{ij}$ **MERW:** $\Delta_{ij} = M_{ij}\frac{\psi_i\psi_j}{\lambda} - \psi_i^2\cdot\delta_{ij}$

Normalized MERW Laplacian: $\left(\Delta_{\mathrm{sym}}\right)_{ij} = \frac{M_{ij}}{\lambda} - \delta_{ij}$

Heat equation and **kernel**: $\frac{d}{dt} K_t = \Delta K_t$ $K_t = \exp(t\Delta) = \lim_{n\to\infty}\left(1 + \frac{t\Delta}{n}\right)^n = \sum_k \frac{(t\Delta)^k}{k!}$

MEPDM – maximal entropy inverse p-distance with matrix exponentiation

Inverse P-distance: $P(i,j) = \sum_{t_{ij}: i \rightsquigarrow j} P[t]\cdot\alpha^{l(t_{ij})}$ (or $\alpha^l/l!$)

for MERW : $l(t_{ij}) = l(t'_{ij}) \Rightarrow P[t_{ij}] = P[t'_{ij}]$ so $P(i,j) = \frac{\psi_j}{\psi_i}\sum_{l\geq 1}\left(\frac{\alpha}{\lambda}\right)^l (A^l)_{ij}$

Hitting/commute time (MFPT): $h(i,j) = [i\neq j](1 + \sum_k S_{ik} h(k,j))$ $c(i,j) = h(i,j) + h(j,i)$

MERW – the most random among random walks

uniform distribution among paths, not edges (GRW)

- As the choice of statistical parameters of an **informational channel**
  MERW allows to maximize channel capacity under some constraints (language?)

- **As random walk/diffusion** (scale-free)
  GRW: the walker indeed performs succeeding random decisions
  MERW: only represents our (lack of) knowledge about a complex dynamics

- **For metrics to analyze complex network**
  GRW sees only degrees of vertices, poorly distinguish nodes
  MERW allows to evaluate importance in the space of possible paths

- social/evolutionary entropy (Lloyd Demetrius):
  "thinking" in terms of paths (reason→result chains) of possibilities?

GRW → MERW

in many cases improves performance or agreement