SFB 680
Molecular Basis of
Evolutionary Innovations

Fitness landscapes and adaptive evolution

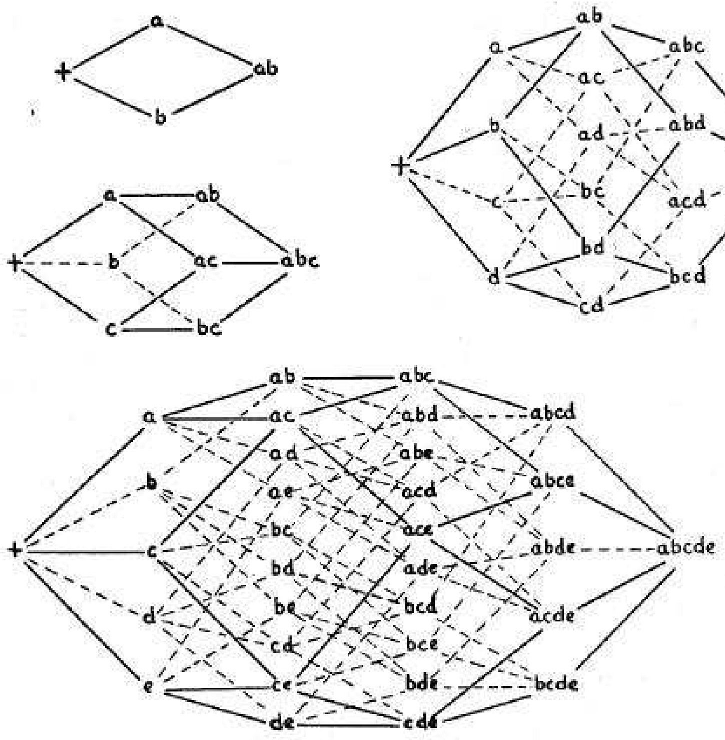
Joachim Krug
Institute for Theoretical Physics, University of Cologne, Germany

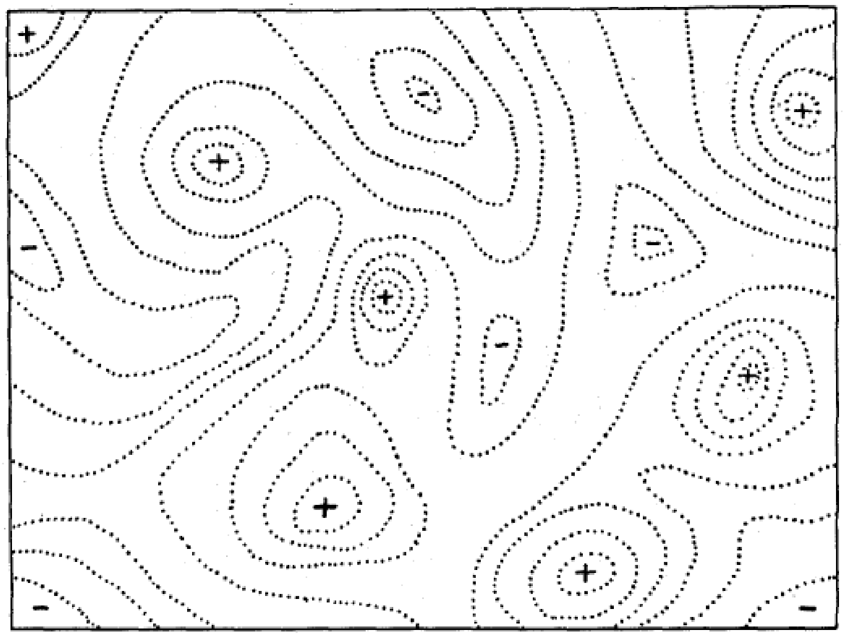
- Empirical fitness landscapes and measures of epistasis
- Accessible mutational pathways in random field models
- Adaptive walks

“Population Dynamics and Statistical Physics in Synergy“
EURANDOM, Eindhoven, August 27, 2014

Fitness landscapes

S. Wright, Proc. 6th Int. Congress of Genetics (1932)





”...selection will easily carry the species to the nearest peak, but there will be innumerable other peaks that will be higher but which are separated by ‘valleys’. The problem of evolution as I see it is that of a mechanism by which the species may continually find its way from lower to higher peaks...”

Mathematical setting

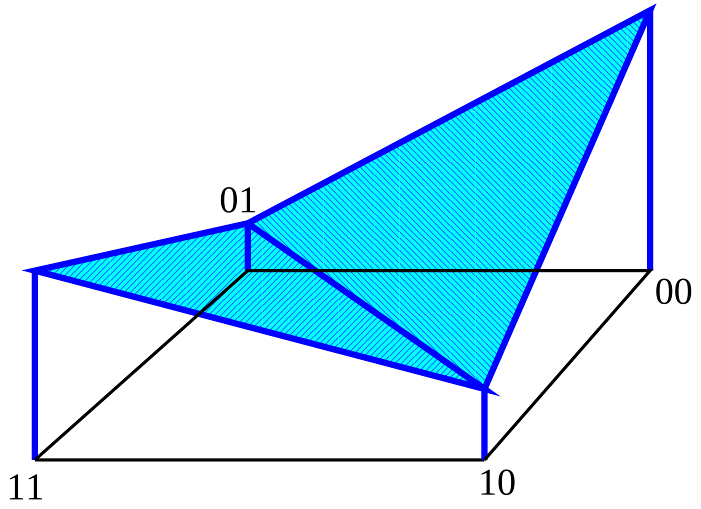
- Genotypes are binary sequences $\sigma = (\sigma_1, \sigma_2, ..., \sigma_L)$ with $\sigma_i \in \{0,1\}$ or $\sigma_i \in \{-1,1\}$ (presence/absence of mutation).

- A fitness landscape is a function $f(\sigma)$ on the space of 2^L genotypes

- Epistasis implies interactions between the effects of different mutations

- Sign epistasis: Mutation at a given locus is beneficial or deleterious depending on the state of other loci Weinreich, Watson & Chao (2005)

- Reciprocal sign epistasis for $L = 2$:



Measures of epistasis

Local fitness optima

Haldane 1931, Wright 1932

- A genotype σ is a local optimum if $f(\sigma) > f(\sigma')$ for all one-mutant neighbors σ'
- In the absence of sign epistasis there is a single global optimum
- Reciprocal sign epistasis is a necessary but not sufficient condition for the existence of multiple fitness peaks Poelwijk et al. 2011, Crona et al. 2013

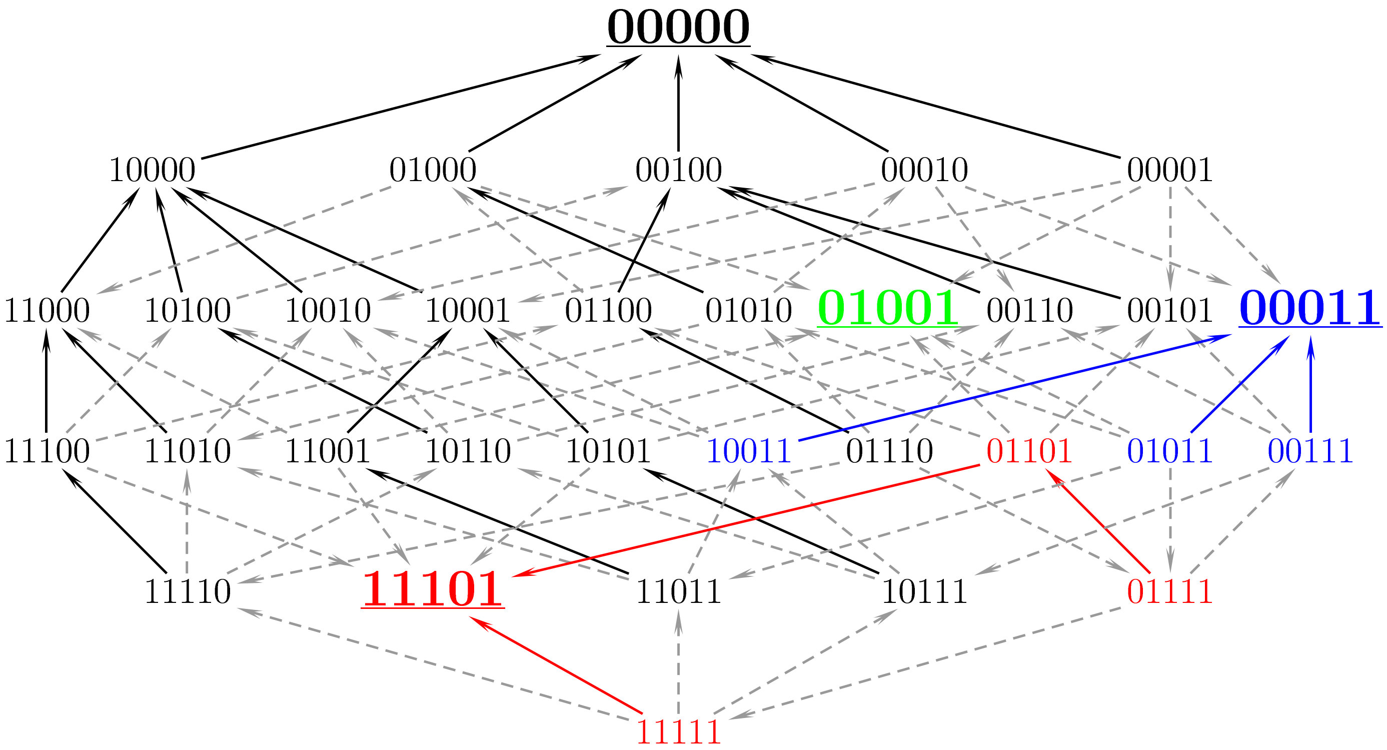
Selectively accessible paths

Weinreich et al. 2005

- A path of single mutations connecting two genotypes $\sigma \longrightarrow \sigma'$ with $f(\sigma) < f(\sigma')$ is selectively accessible if fitness increases monotonically along the path
- In the absence of sign epistasis all paths to the global optimum are accessible, and vice versa

Empirical example: The *Aspergillus niger* fitness landscape

J.A.G.M. de Visser, S.C. Park, JK, American Naturalist **174**, S15 (2009)



- Combinations of 8 individually deleterious marker mutations (one out of $\binom{8}{5} = 56$ five-dimensional subsets shown)

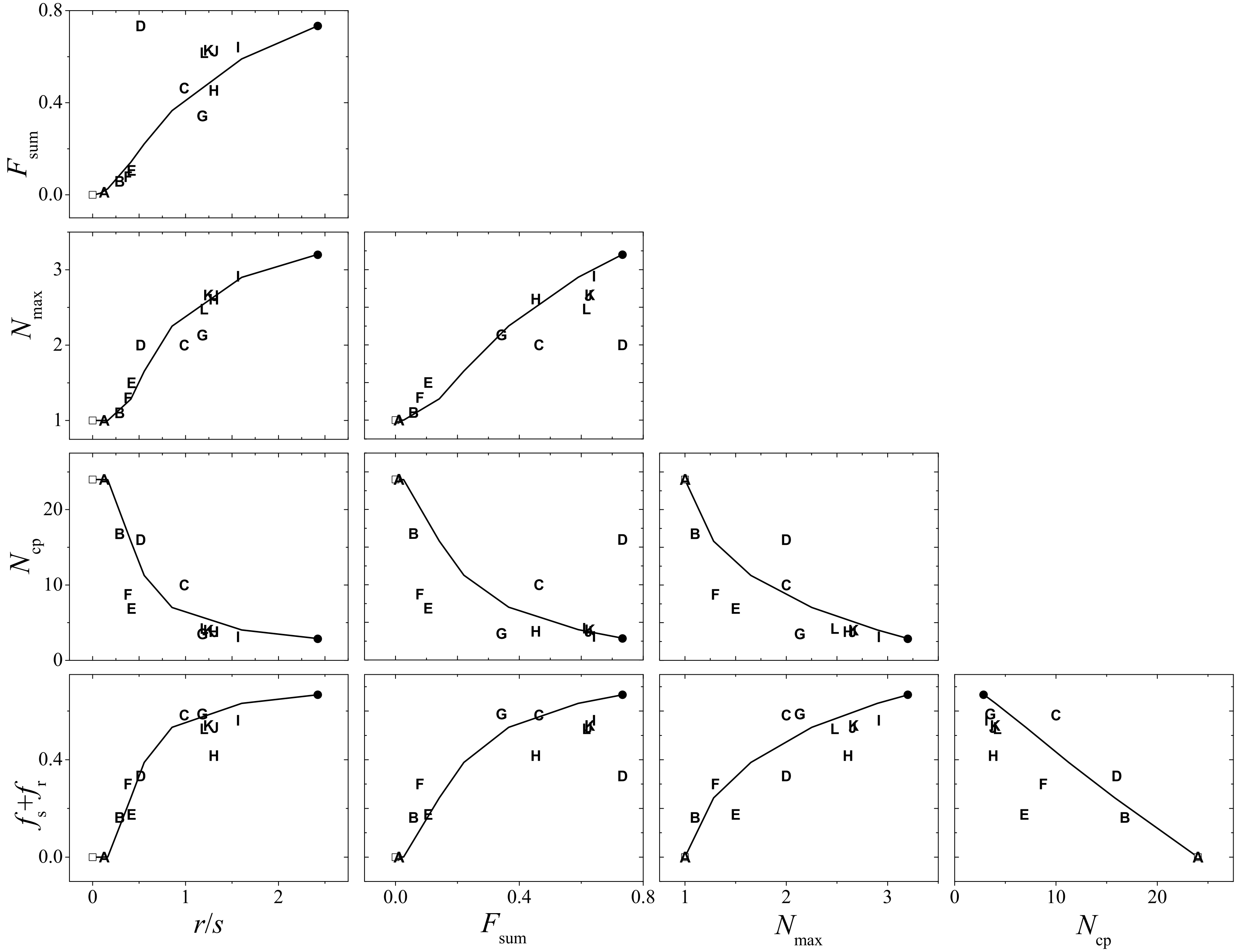
- 3 local fitness optima, 25 out of 120 paths are accessible

A metaanalysis of empirical data sets

Szendro et al., JSTAT 2013; de Visser & Krug, Nat. Rev. Genet. 2014

ID	System (*organism*/gene)	L	Available combinations	Fitness (proxy)	Direction of mutations	Known effects
A	*Methylobacterium extorquens*	4	16/16	Growth rate	Beneficial	Combined
B	*Escherichia coli*	5	32/32	Fitness	Beneficial	Combined
C-D	Dihydrofolate reductase	4	16/16	Resistance/ Growth rate	Beneficial	Individual/ Combined
E	β-lactamase	5	32/32	Resistance	Beneficial	Combined
F	β-lactamase	5	32/32	Resistance	Beneficial	Combined
G	*Saccharomyces cerevisiae*	6	64/64	Growth rate	Deleterious	Individual
H	*Aspergillus niger*	8	186/256	Growth rate	Deleterious	Individual
I-J	Terpene synthase	9	418/512	Enzymatic specificity	–	–

Comparison of epistasis measures



Random field models of fitness landscapes

Null model: House-of-cards

- In the house-of-cards model fitness is assigned randomly to genotypes

  Kingman 1978, Kauffman & Levin 1987

- What is the expected number of shortest, selectively accessible paths n_{acc} from an arbitrary genotype at distance d to the global optimum?

- The total number of paths is $d!$, and a given path consists of d independent, identically distributed fitness values $f_0, \ldots., f_{d-1}$.

- A path is accessible iff $f_0 < f_1 \ldots. < f_{d-1}$

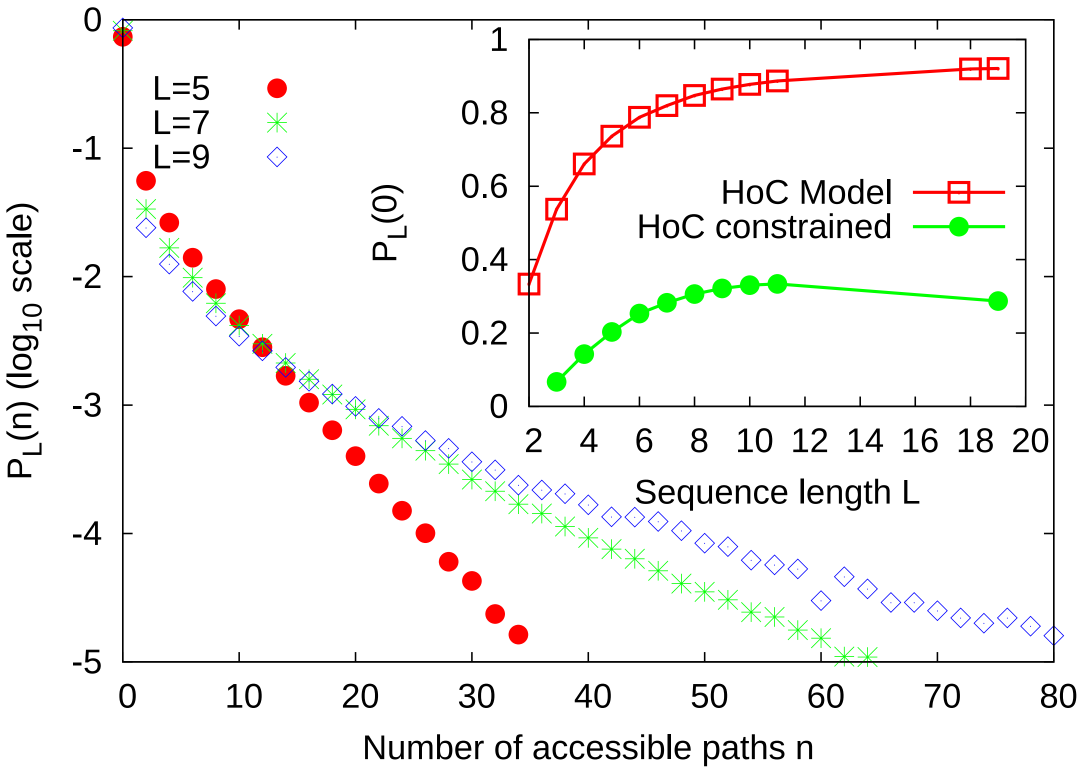
- Since all $d!$ permutations of the d random variables are equally likely, the probability for this event is $1/d!$

$$\Rightarrow \ \mathbb{E}(n_{\text{acc}}) = \frac{1}{d!} \times d! = 1$$

- This holds in particular for the $L!$ paths from the reversal genotype/antipode of the global optimum.

Distribution of number of accessible paths from reversal genotype

J. Franke et al., PLoS Comp. Biol. **7** (2011) e1002134



- "Condensation of probability" at $n_{\mathrm{acc}} = 0$

- Characterize the distribution $P_L(n)$ by $\mathbb{E}(n_{\mathrm{acc}})$ and the probability $P_L(0)$ that no path is accessible ⇒ define accessibility as $\overline{P}_L \equiv 1 - P_L(0)$

“Accessibility percolation” as a function of initial fitness

- When fitnesses are drawn from the uniform distribution and the fitness of the initial genotype is f_0, then Hegarty & Martinsson, Ann. Appl. Prob. 2014

$$\lim_{L \to \infty} \overline{P}_L = \begin{cases} 0 \quad \text{for} \quad f_0 > \dfrac{\ln L}{L} \\[2ex] 1 \quad \text{for} \quad f_0 < \dfrac{\ln L}{L}, \end{cases}$$

- This implies in particular that $\lim_{L \to \infty} \overline{P}_L = 0$ for the HoC model with unconstrained initial fitness
- If arbitrary paths with backsteps are allowed, the accessibility threshold becomes independent of L Berestycki, Brunet, Shi, arXiv:1401.6894
- On a regular tree of height h and branching number b the accessibility threshold for $h, b \to \infty$ occurs at $h/b = e$

Nowak & Krug, EPL 2013; Roberts & Zhao, ECP 2013

Landscapes with tunable ruggedness

Kauffman's NK-model

Kauffman & Weinberger 1989

- Each locus interacts randomly with $K \leq L-1$ other loci:

$$f(\sigma) = \sum_{i=1}^{L} f_i(\sigma_i|\sigma_{i_1}, \ldots, \sigma_{i_K})$$

f_i: Uncorrelated RV's assigned to each of the 2^{K+1} possible arguments

- $K = 0$: Non-epistatic $K = L-1$: House-of-cards

Rough Mt. Fuji model

Aita et al. 2000; Neidhart et al., arXiv:1402.3065

- Non-epistatic ("Mt. Fuji") landscape perturbed by a random component:

$$f(\sigma) = -cd(\sigma, \sigma^{(0)}) + \eta(\sigma)$$

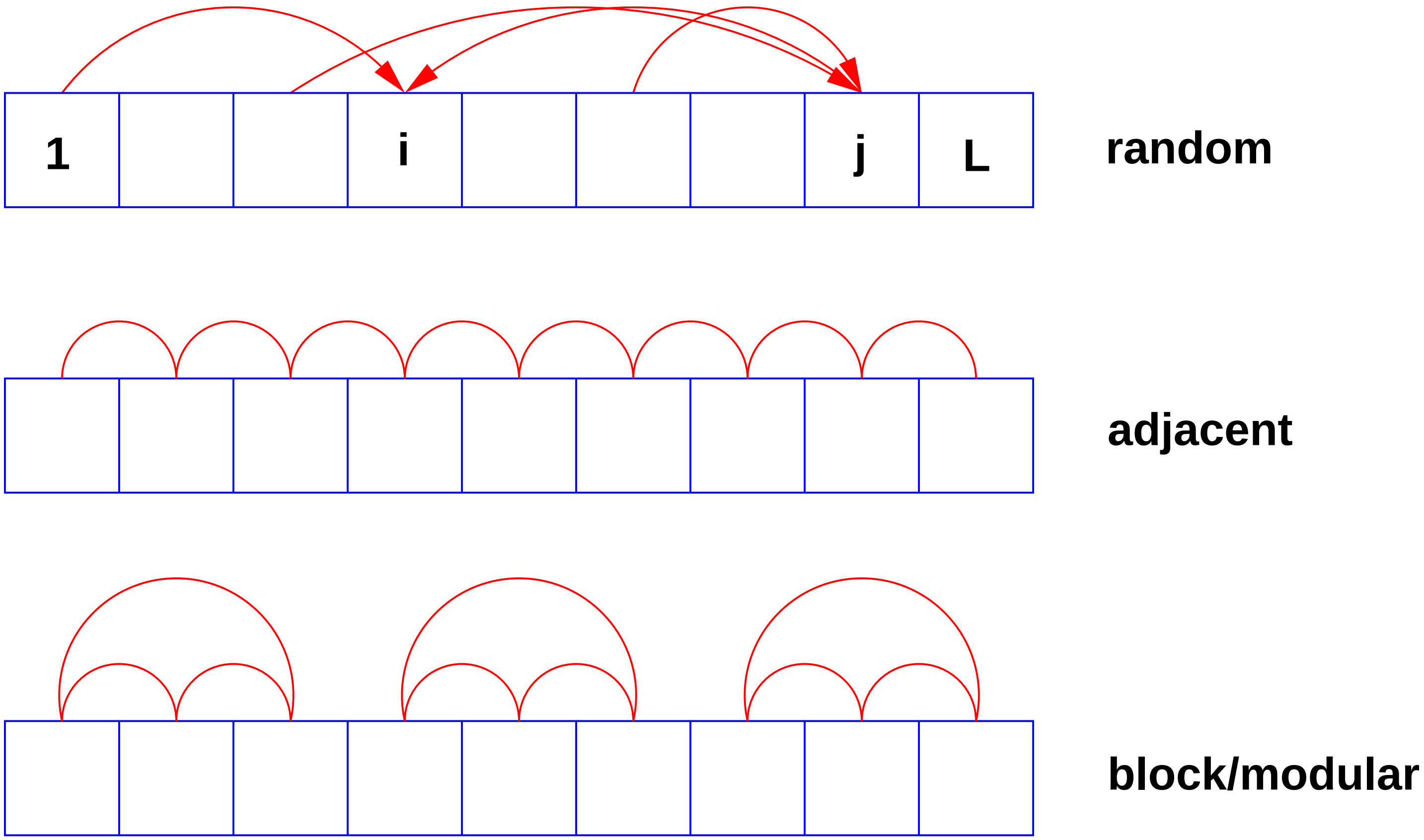
$c > 0$: slope $d(\sigma, \sigma')$: Hamming distance η: i.i.d. RV's

- $\lim_{L\to\infty} \overline{P}_L = 1$ for any $c > 0$

Hegarty & Martinsson 2014

“Genetic architecture” in Kauffman’s NK-model

- Different schemes for choosing the interaction partners:



- Which properties of the fitness landscape are sensitive to this choice?

“Genetic architecture” in Kauffman’s NK-model

- Fitness correlation function is manifestly independent of the neighborhood scheme

  P.R.A. Campos, C. Adami, C.O. Wilke (2002)

  J. Neidhart, I.G. Szendro, JK, JTB 2013

- In the block model, the mean number of local maxima is given exactly by

  $$\mathbb{E}(n_{\max}^{\text{block}}) = \frac{2^L}{(K+2)^{L/(K+1)}}$$

  A.S. Perelson, C.A. Macken (1995)

  which is very close (but not identical) to rigorous results for the adjacent model

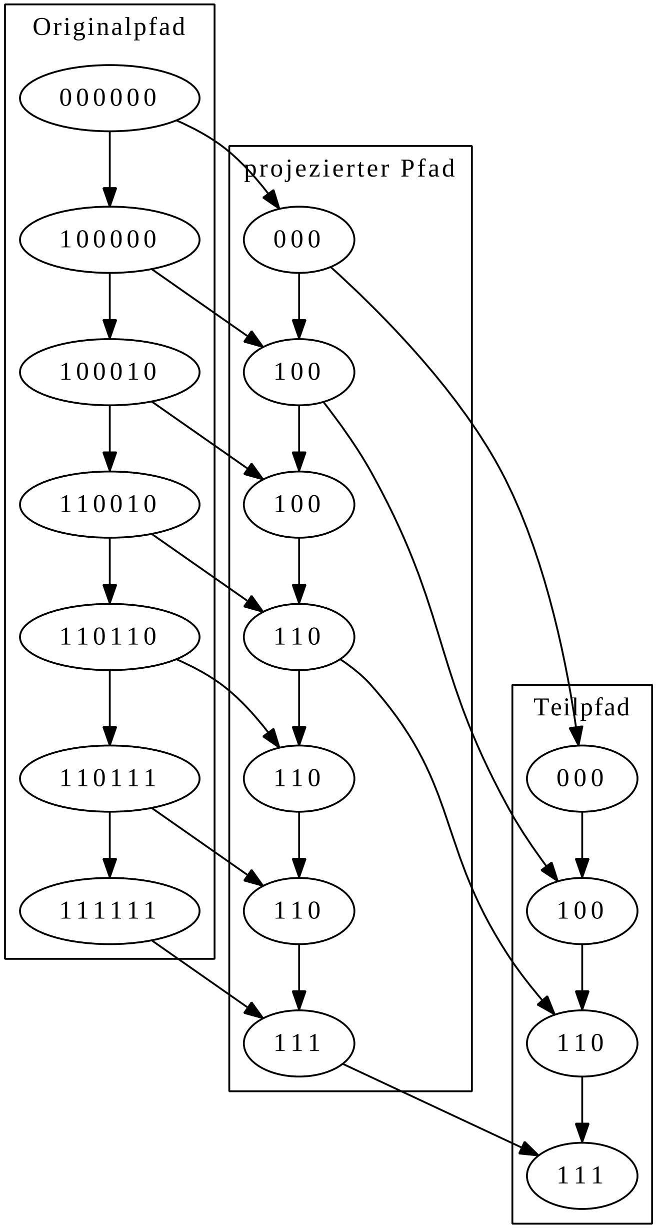
  Durrett & Limic (2003), Limic & Pemantle (2004)

- Mean number of accessible paths in the block model:

  $$\mathbb{E}(n_{\text{acc}}^{\text{block}}) = \frac{L!}{[(K+1)!]^{L/(K+1)}}$$

  B. Schmiegelt, JK 2013

Path decomposition for the block model



Evolutionary accessibility in the block model

B. Schmiegelt, JK, J. Stat. Phys. **154**, 334 (2014)

- A given pathway spanning the whole landscape is accessible iff all subpaths within the $B = L/(K+1)$ blocks are accessible

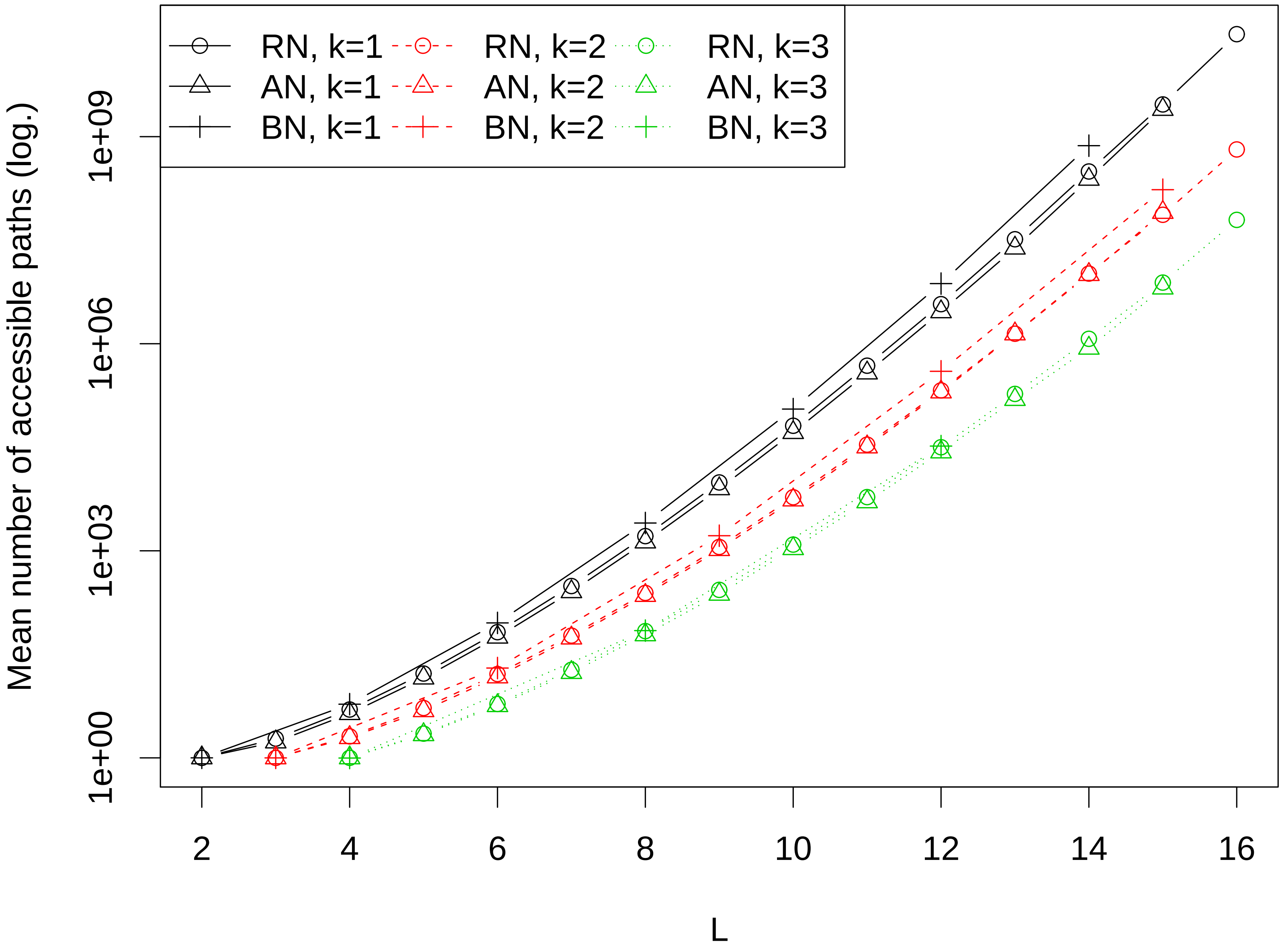
- Each combination of accessible subpaths can be combined into $\frac{L!}{[(K+1)!]^B}$ global paths

$$\Rightarrow \quad n_{\text{acc}}^{\text{block}} = \frac{L!}{[(K+1)!]^B} \prod_{i=1}^{B} n_{\text{acc}}^{(i)}$$

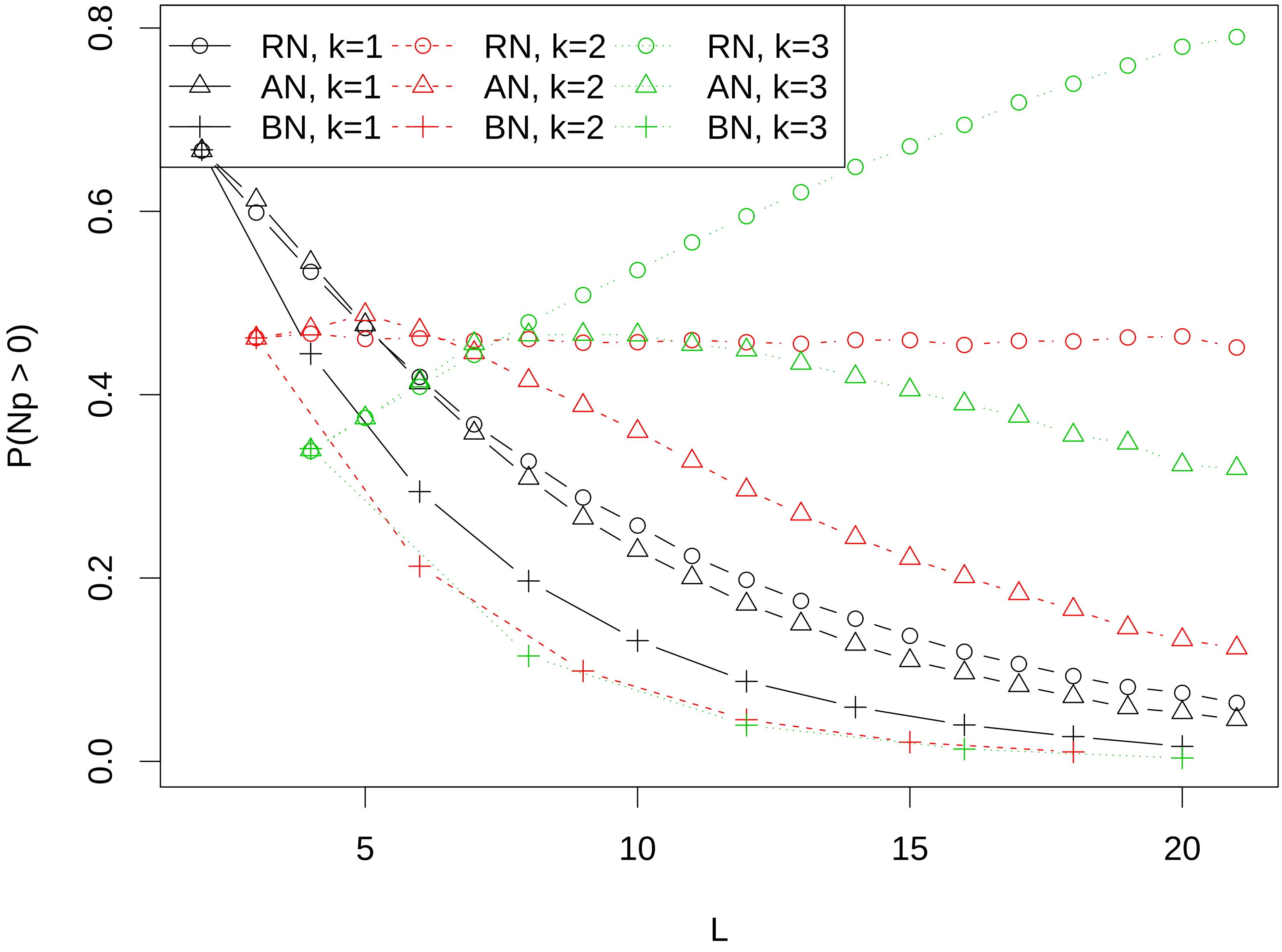
- Since the blocks are HoC-landscapes of size $K+1$, the expected number of accessible paths is $\mathbb{E}(n_{\text{acc}}^{\text{block}}) = \frac{L!}{[(K+1)!]^B}$ and the accessibility is $\overline{P}_L^{\text{block}} = [\overline{P}_{K+1}^{\text{HoC}}]^{\frac{L}{K+1}}$ which approaches zero exponentially fast in L for any K

- Full distribution of $n_{\text{acc}}^{\text{block}}$ can be computed in terms of the HoC distributions, explicit results for $K = 1$ and $K = 2$.

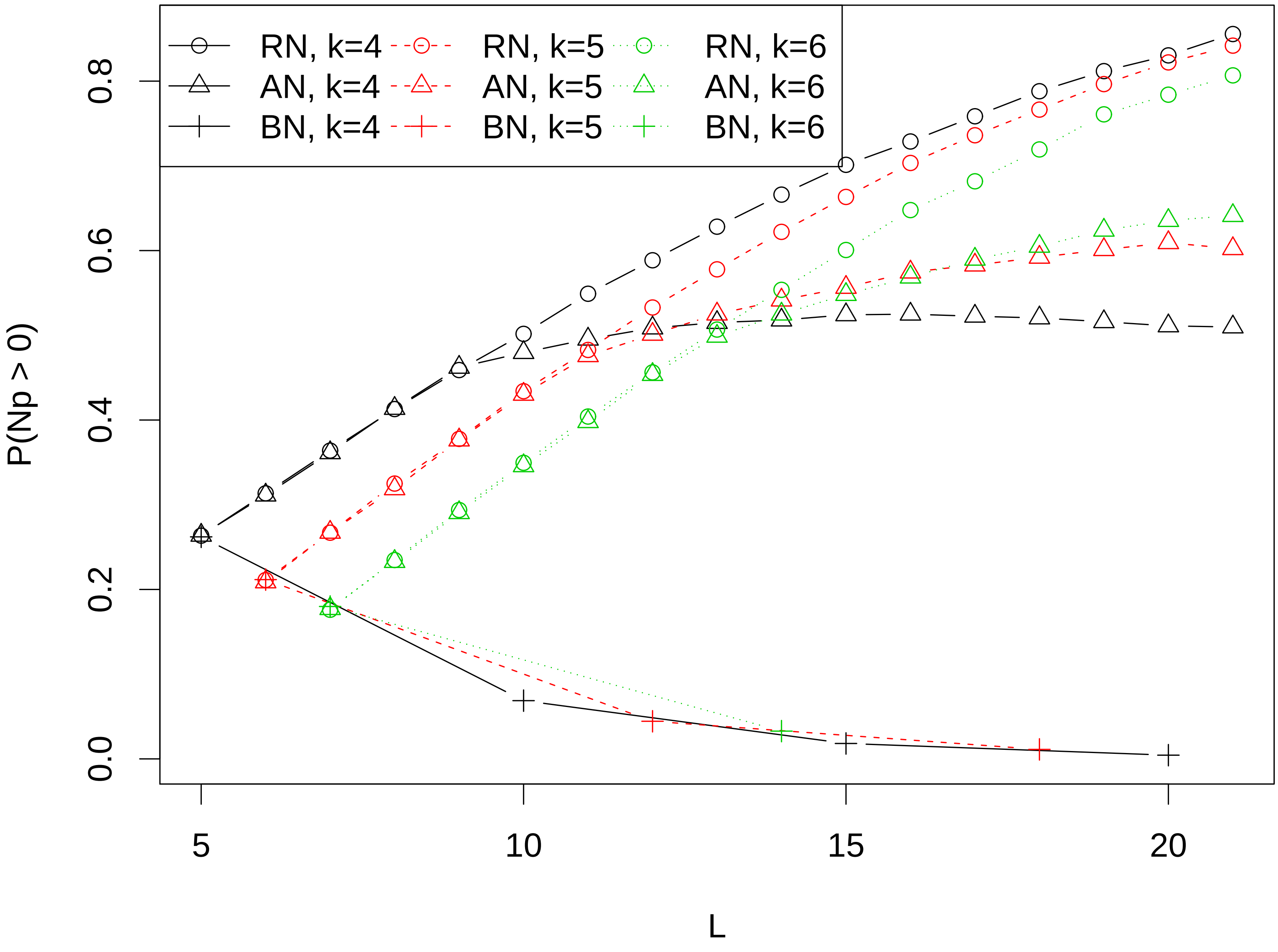
Mean number of paths is insensitive to genetic architecture



...but accessibility is very sensitive....

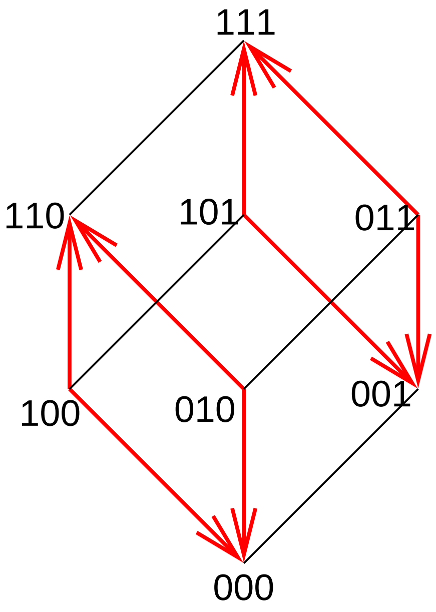


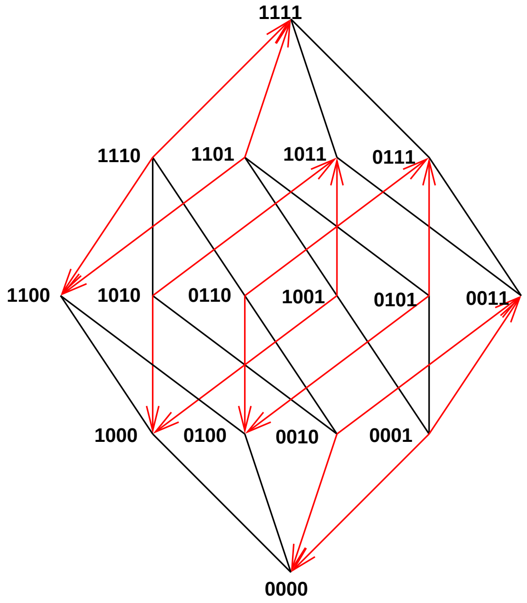
...at least for system sizes that can be simulated!



Global reciprocal sign epistasis

- A fitness landscape displays global reciprocal sign epistasis if there is a pair of loci that has reciprocal sign epistasis in all posssible backgrounds:

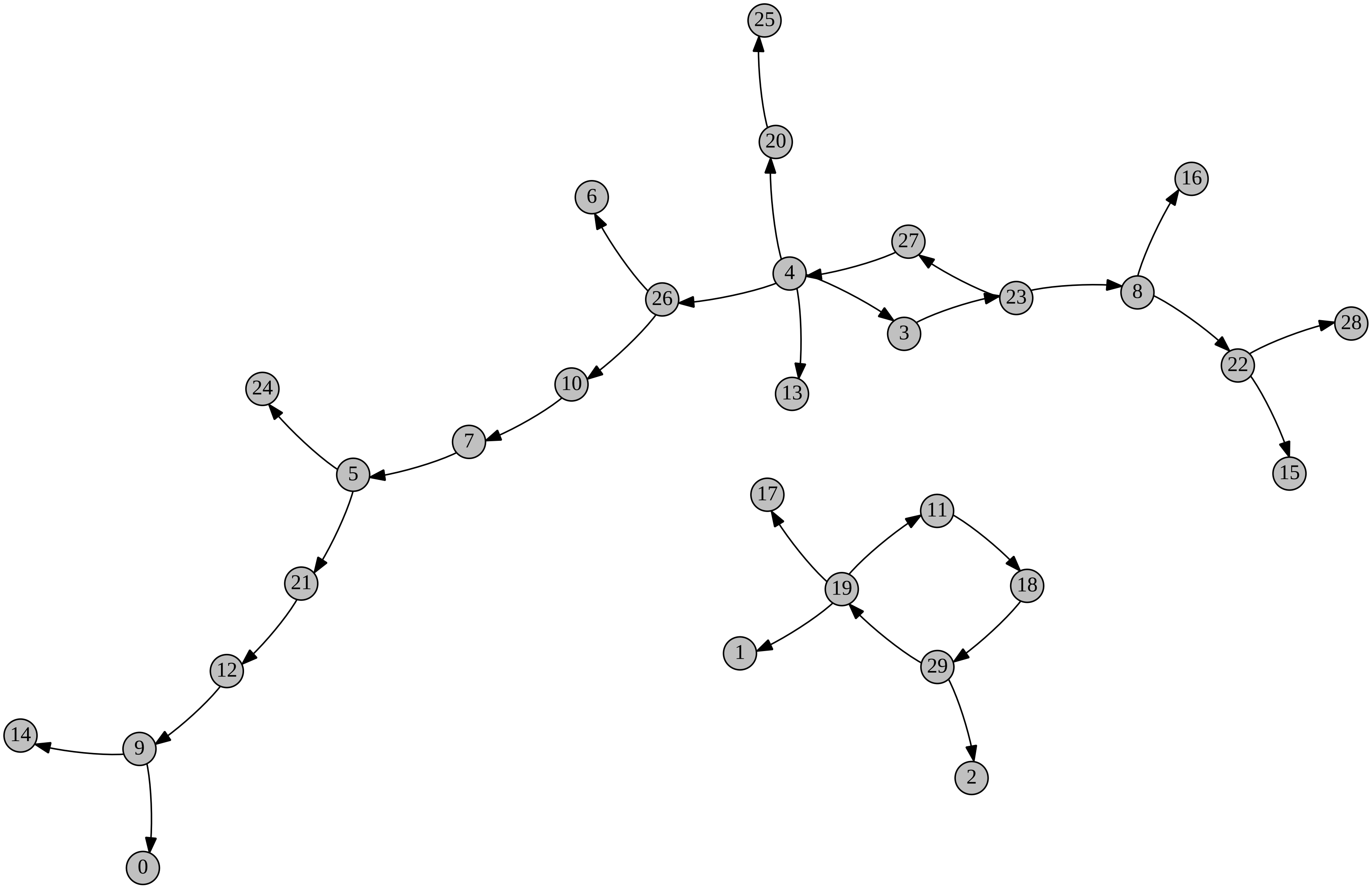




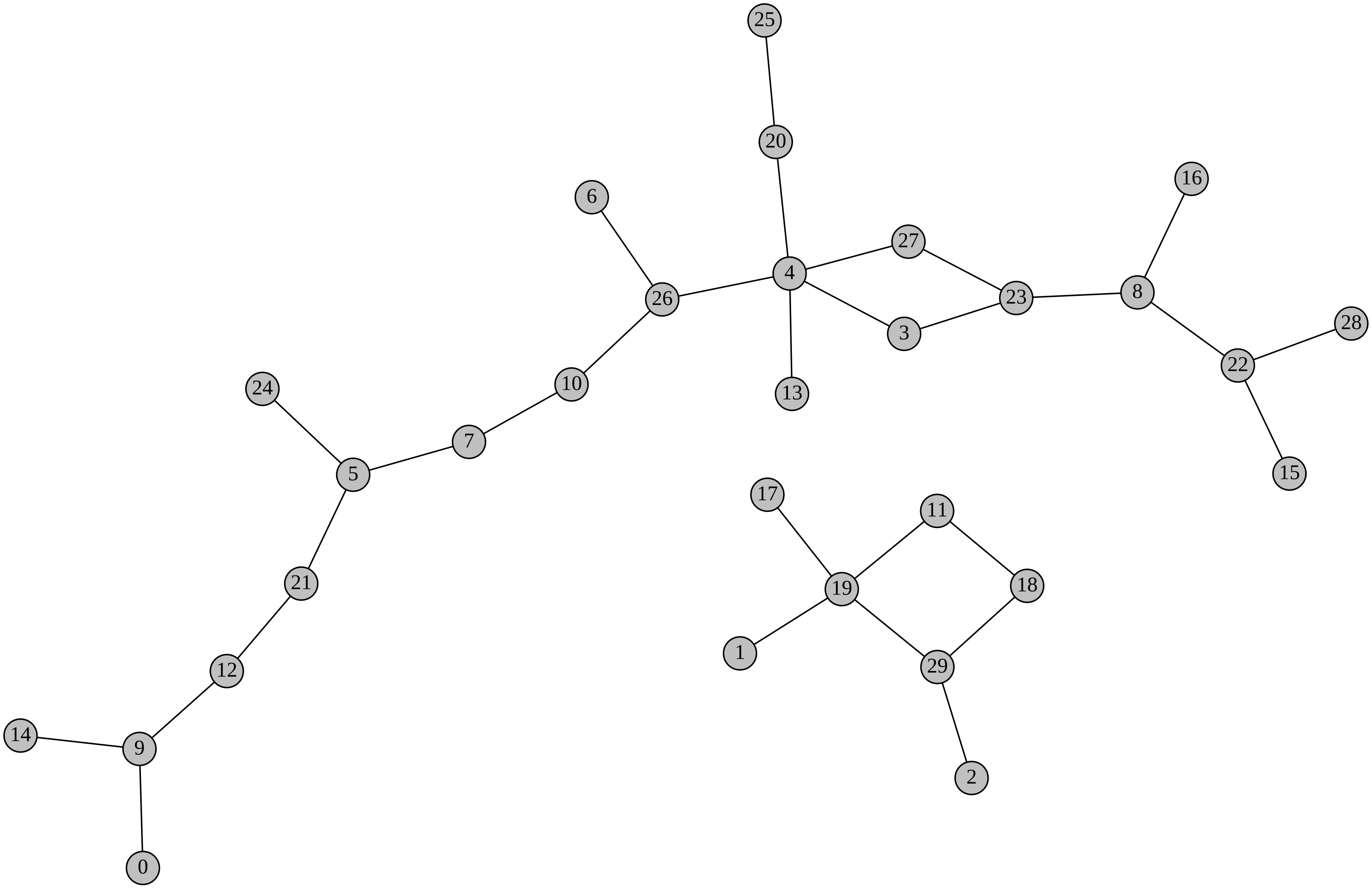
- In the presence of global reciprocal sign epistasis there are no accessible paths across the hypercube

- **Proposition:** (B. Schmiegelt)
  Global reciprocal sign epistasis exists with probability tending to unity for $L \longrightarrow \infty$, K fixed, for any neighborhood choice of the NK-model

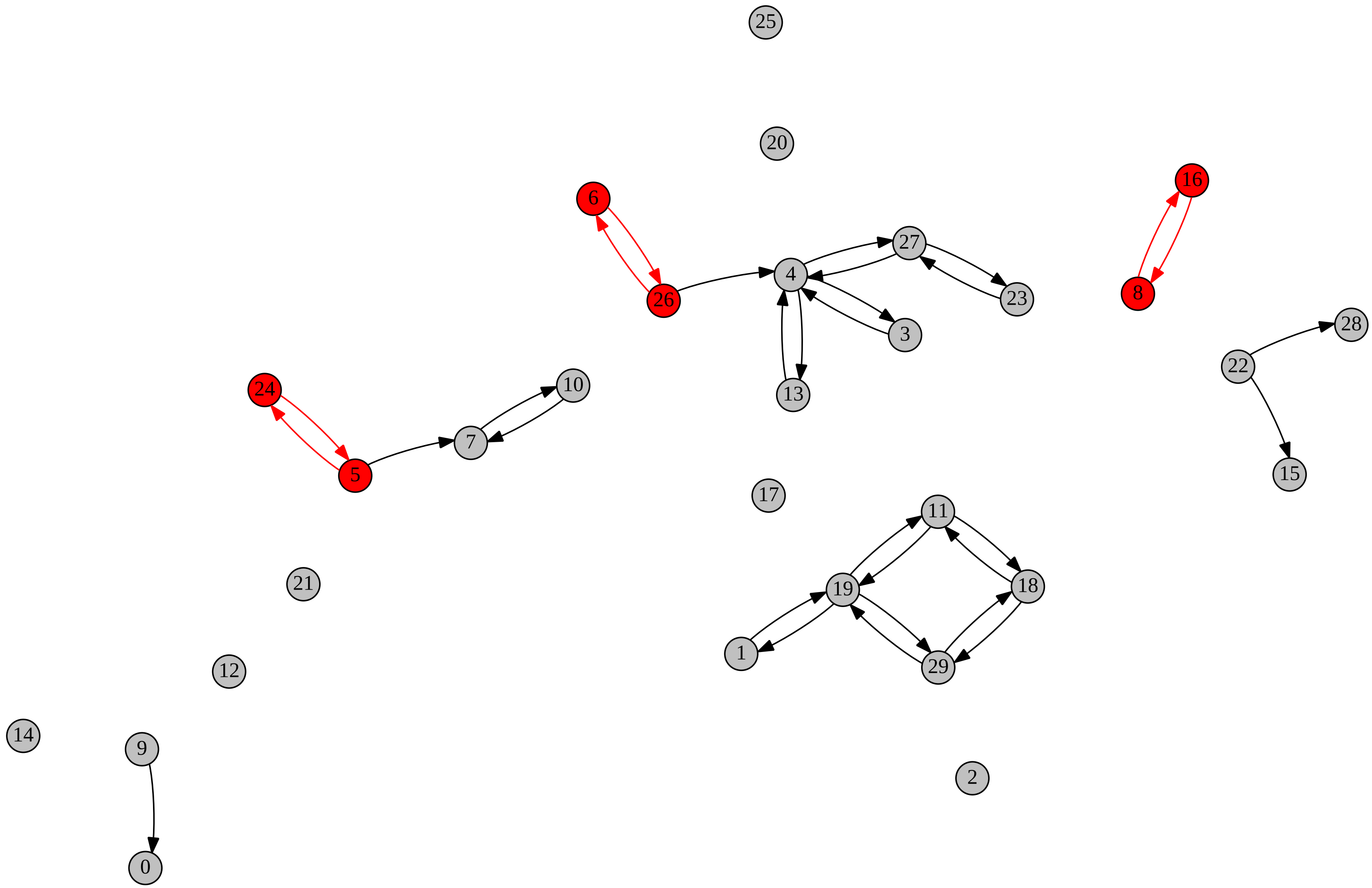
Random NK-model with $L = 30, K = 1$: Neighborhood graph



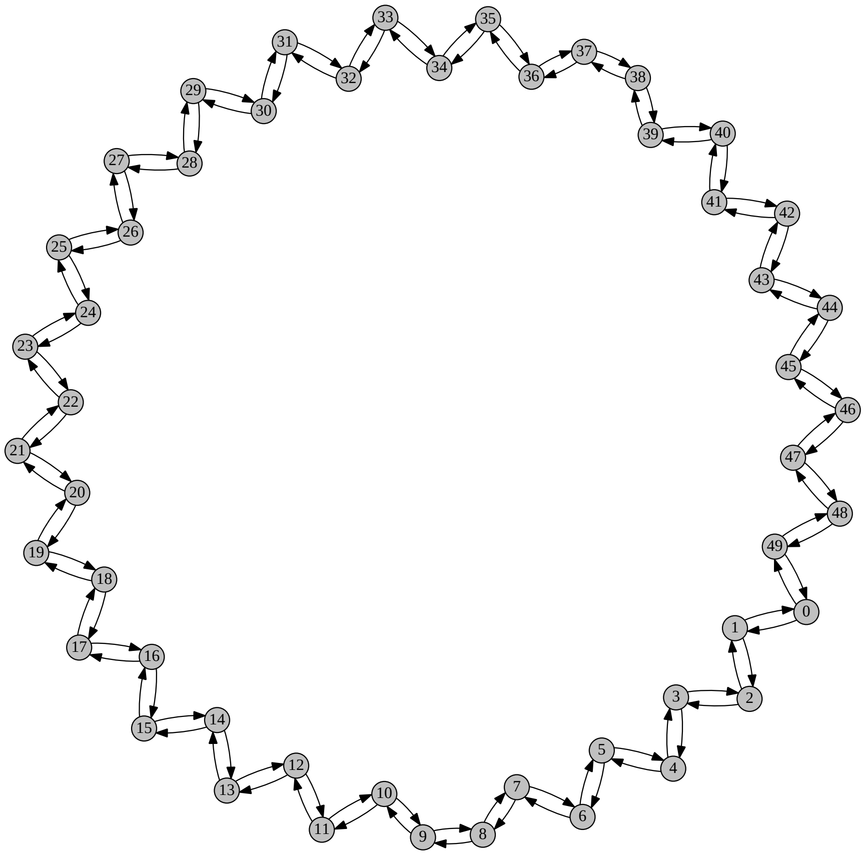
Random NK-model with $L = 30, K = 1$: Epistasis graph

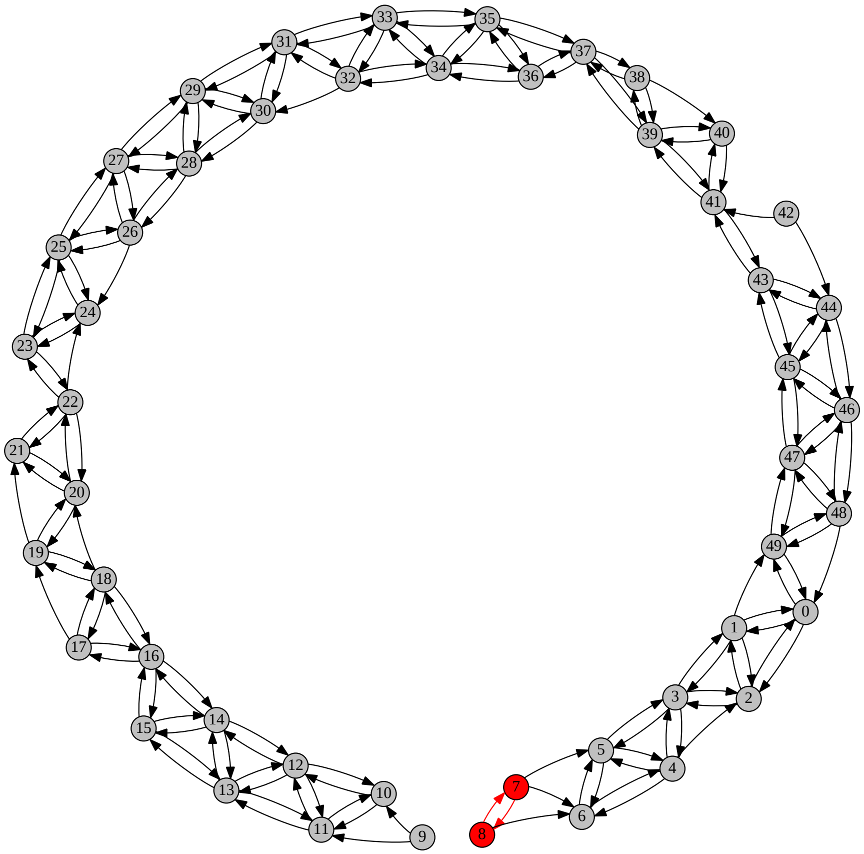


Random NK-model with $L = 30, K = 1$: Sign epistasis graph



Adjacent NK-model with $L = 50, K = 2$

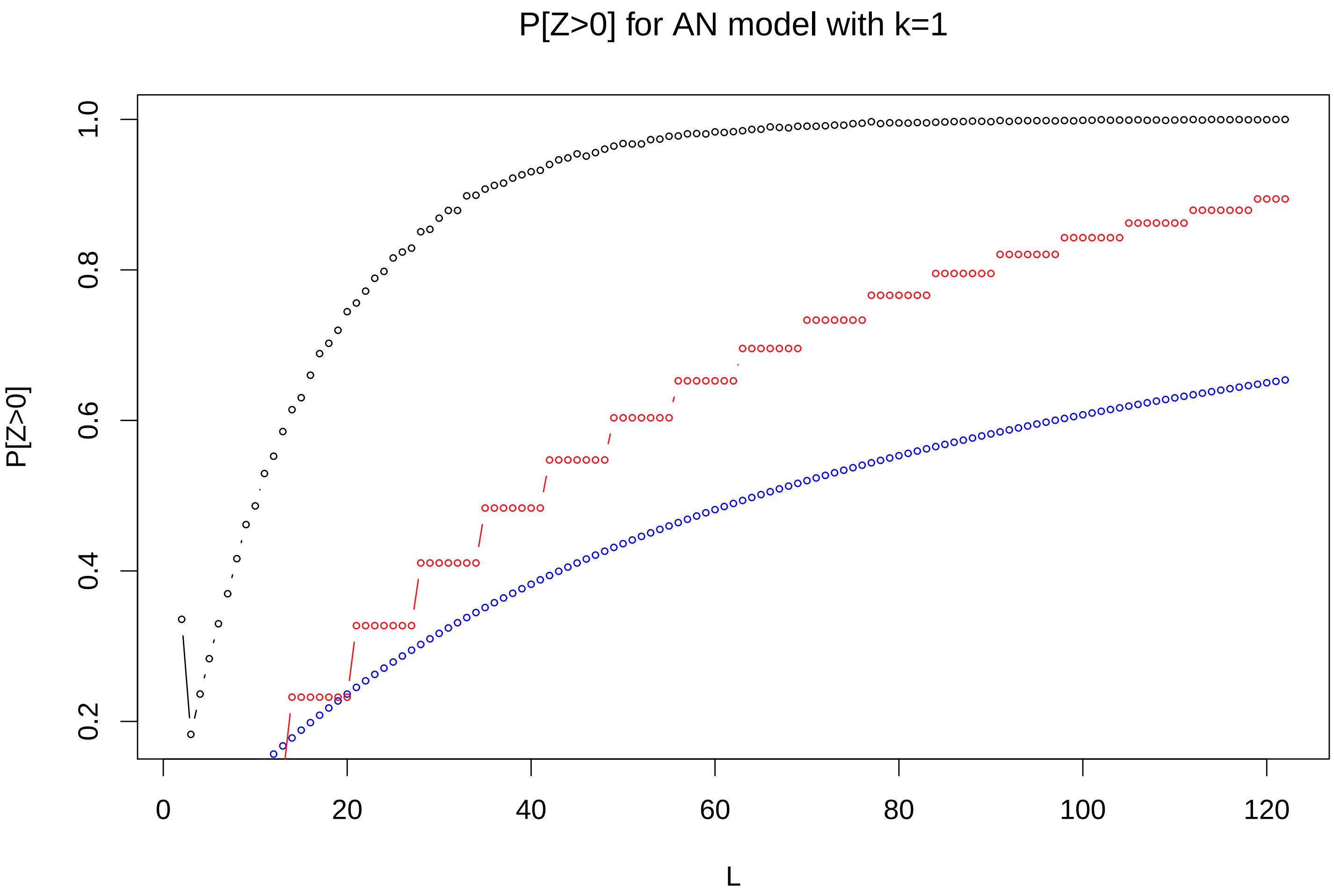




- ● interaction graph
- ● sign epistasis graph

Global reciprocal sign epistasis in the adjacent NK-model: $K = 1$

B. Schmiegelt (unpublished)

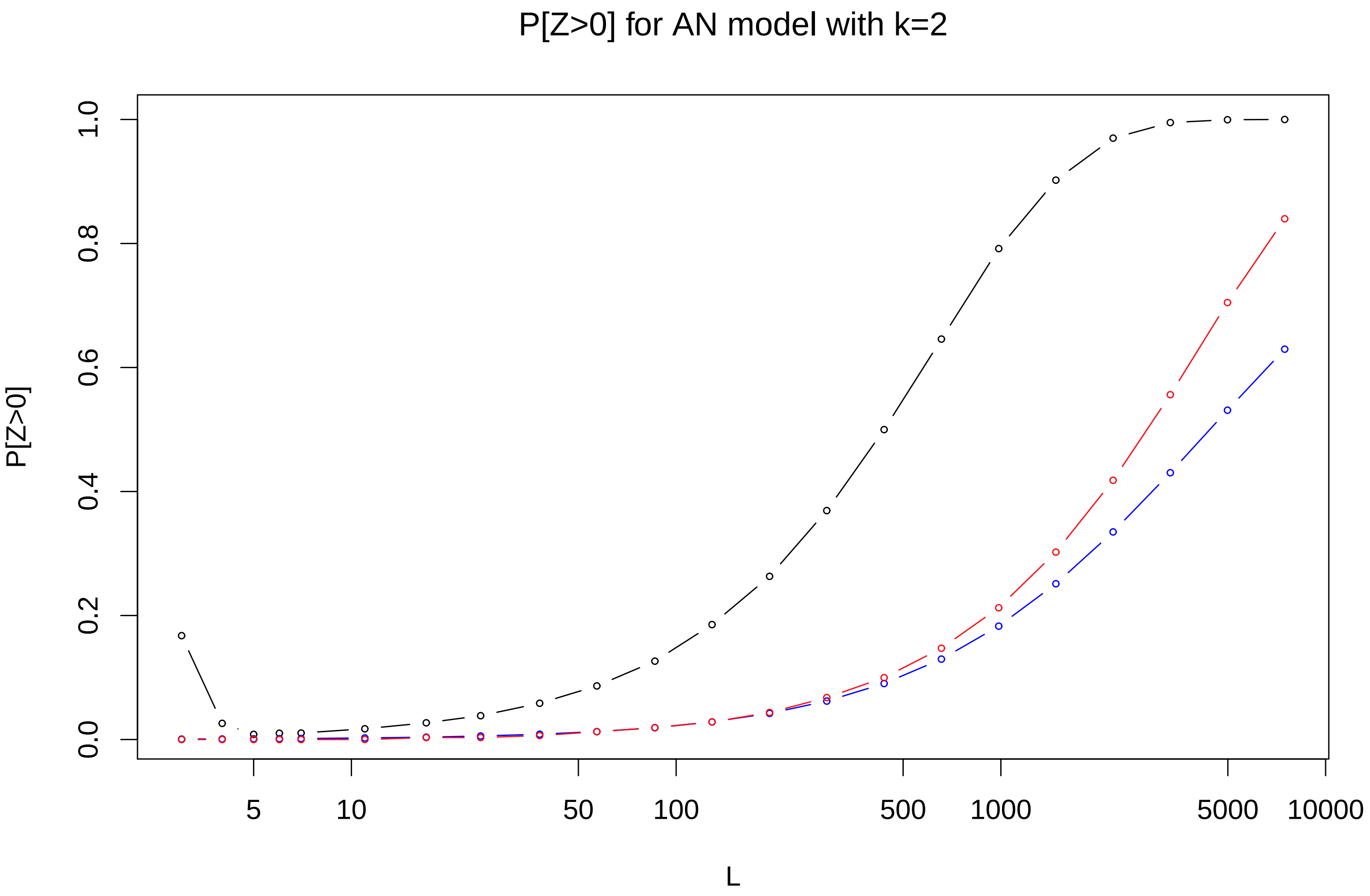


black: numerical

red and blue: analytic bounds

Global reciprocal sign epistasis in the adjacent NK-model: $K = 2$

B. Schmiegelt (unpublished)



black: numerical

red and blue: analytic bounds

Adaptive walks

Adaptive walks

- An adaptive walk is a Markov chain on sequence space that is constrained to move to genotypes of larger fitness and terminates at local fitness maxima

- Three flavors of adaptive walks differing in their transition probabilities:

  Random Adaptive Walk (RAW) Macken & Perelson 1989
  All fitter genotypes are chosen with equal probability

  Greedy Adaptive Walks (GAW) Orr 2003
  The most fit genotype is chosen deterministically

  True Adaptive Walk (TAW)
  Transition rate is proportional to the fitness difference between the resident and mutant genotype Gillespie 1983, Orr 2002

- Quantities of interest: Average length ℓ and achieved fitness (height) f^*

Walk length in the HoC landscape

- RAW's and GAW's are fully determined by the rank ordering of the fitness landscape. Their properties are independent of the fitness distribution and only depend on the number of uphill directions L in the initial state.

- RAW: $\ell \approx \ln(L) + 1.1$ for large L Flyvbjerg & Lautrup 1992

- GAW: $\ell \to e - 1 \approx 1.71828...$ Orr 2003

- TAW length asymptotics depends on the extreme value index κ of the fitness distribution according to Neidhart & Krug 2011, Jain 2011

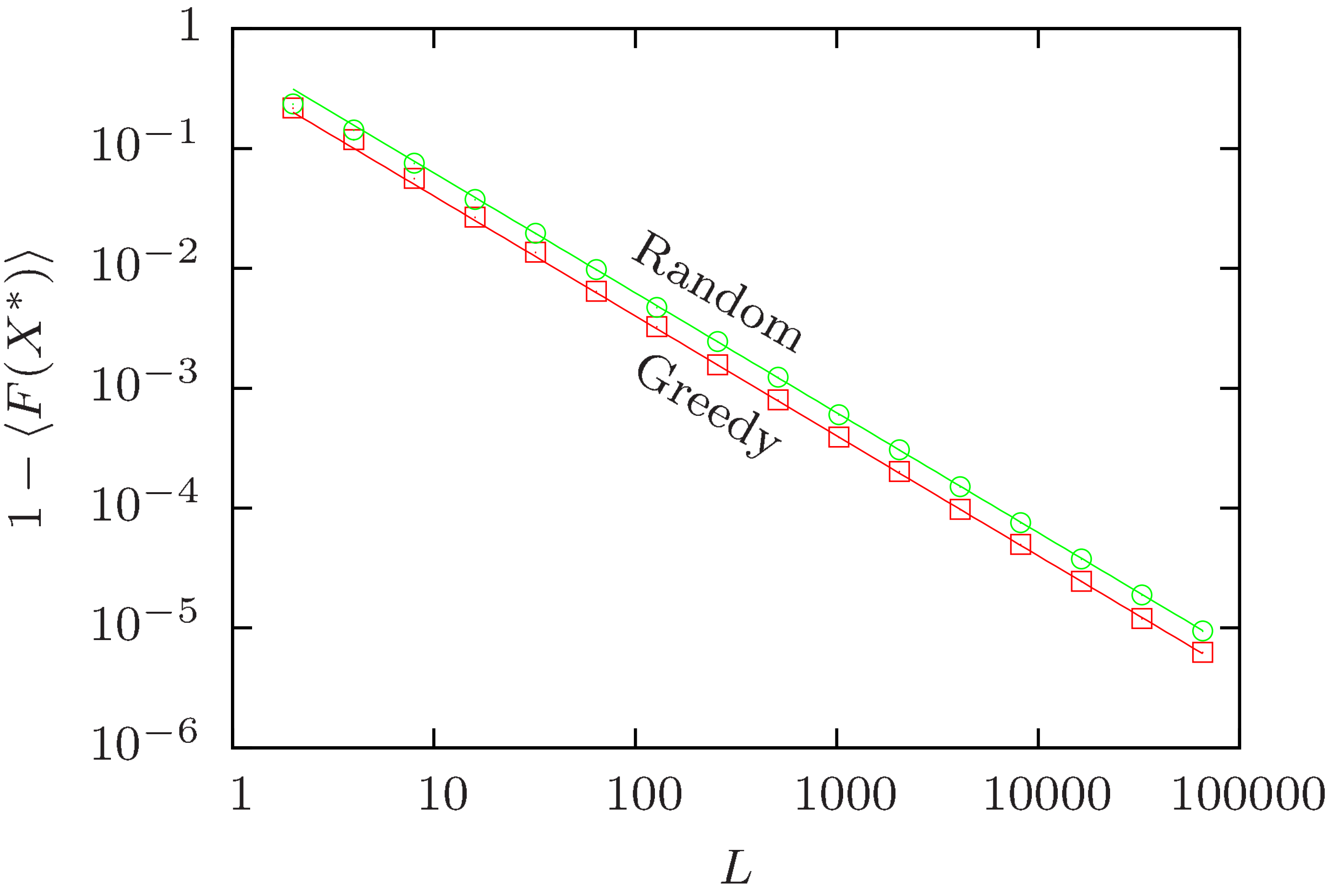
$$\ell \approx \frac{1-\kappa}{2-\kappa} \ln(L) + c_\kappa \quad \text{for} \quad \kappa < 1$$

  where $\kappa > 0$, $\kappa = 0$ and $\kappa < 0$ correspond to the Fréchet, Gumbel and Weibull classes, respectively.

- The TAW becomes effectively random (greedy) for $\kappa \to -\infty$ ($\kappa \to 1$)

Walk height in the HoC landscape

S. Nowak (unpublished)



- For uniform fitness distribution the expected final fitness is of the form $1 - \mathbb{E}(f^*) \approx \frac{\beta}{L}$ with $\beta_{\mathrm{RAW}} \approx 0.6243..$ and $\beta_{\mathrm{GAW}} \approx 0.4003...$

Random adaptive walks in the RMF landscape

S.-C. Park, I.G. Szendro, J. Neidhart, JK, arXiv:1408:4856

- RAW's starting at antipode (maximal distance L from reference sequence)

- Assume RAW takes only 'uphill' steps that decrease $d(\sigma, \sigma^{(0)})$, and draw random fitness component from exponential distribution with mean 1

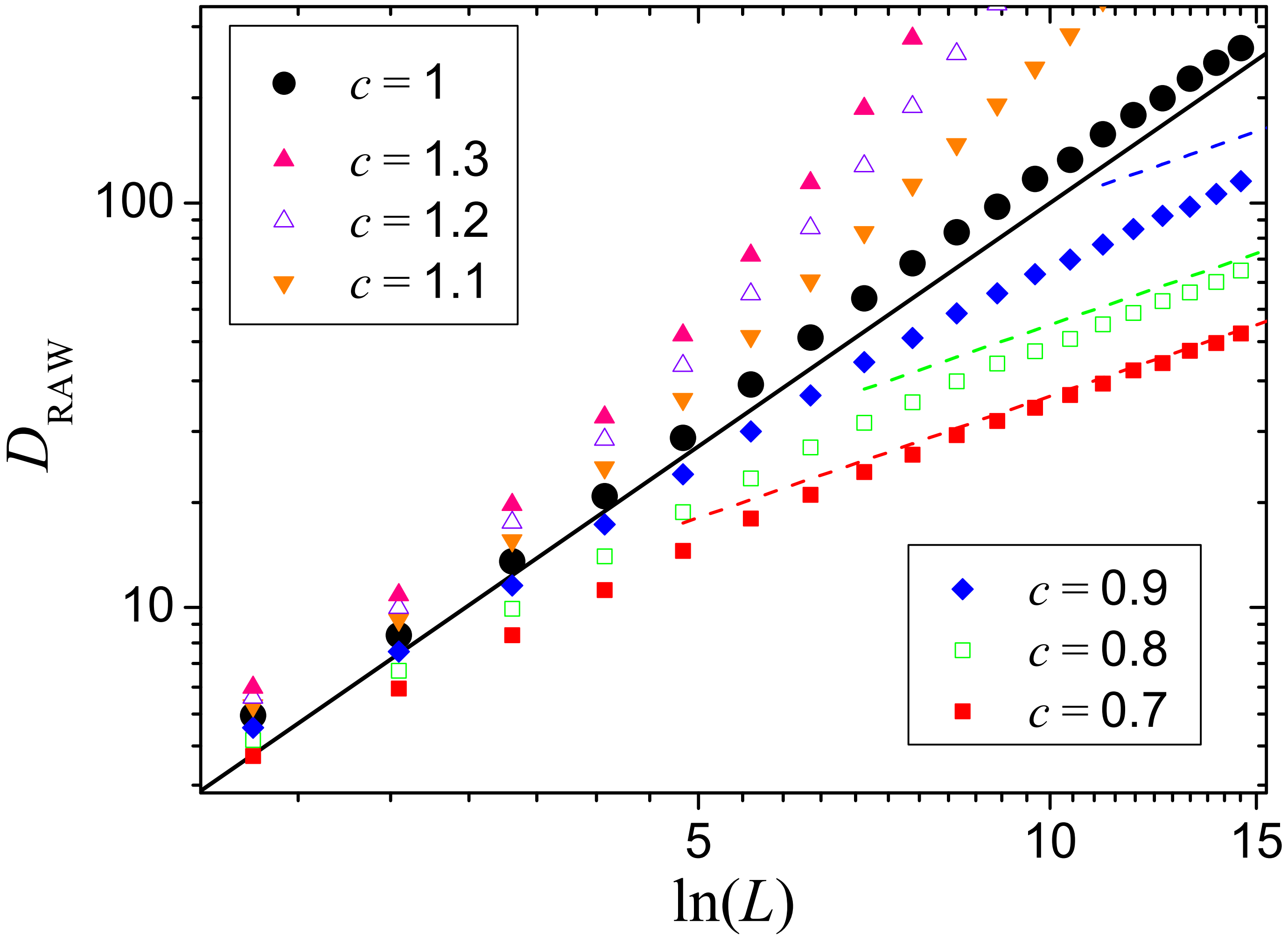
- Then the mean walk length can be computed analytically and displays a phase transition at $c = 1$:

$$\ell \propto \begin{cases} \ln L/(1-c), & c < 1 \\ (\ln L)^2, & c = 1, \\ O(L), & c > 1. \end{cases}$$

- For tails thinner (fatter) than exponential, $\ell \sim L$ ($\ell \sim \ln L$) for all $c > 0$

- Equivalent to zero temperature Metropolis dynamics of a random energy spin glass in an external field

Random adaptive walks in the RMF landscape



- Numerical verification of phase transition in simulations with backsteps

Greedy adaptive walks in the RMF landscape

S.-C. Park, J. Neidhart, JK, in preparation

- For Gumbel-distributed random fitness components, the length of GAW's starting from the antipode of the reference sequence satisfies

$$\mathbb{P}(\mathrm{length} \geq l) = \prod_{k=1}^{l} \frac{1-e^{-c}}{1-e^{-kc}} = \frac{1}{[l]_{e^{-c}}!}$$

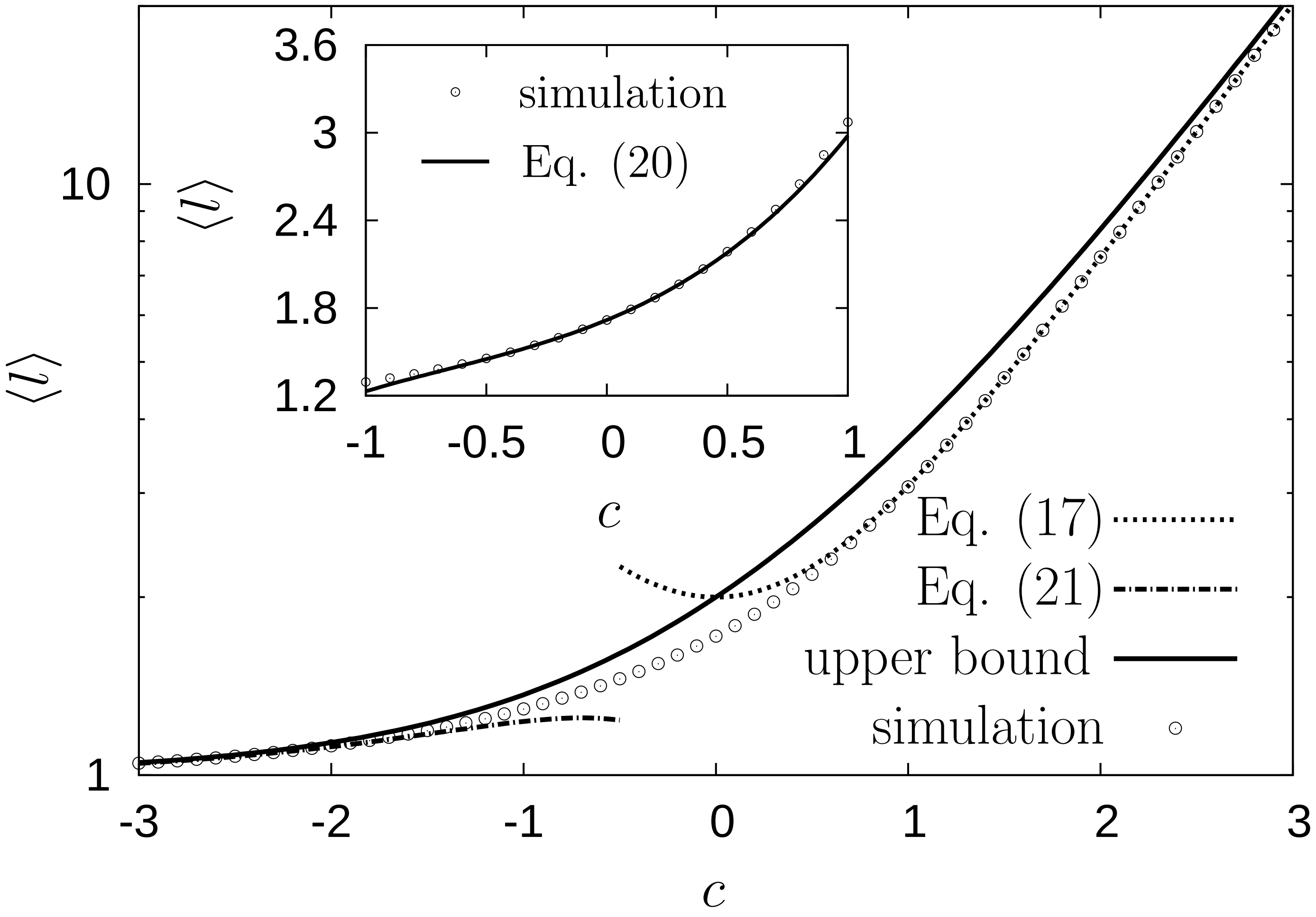
  where $[n]_q = \frac{1-q^n}{1-q}$ is the q-number.

- Correspondingly the mean walk length is given by the q-exponential

$$\ell = \exp_{e^{-c}}(1) - 1 \to e - 1 \quad \text{for} \quad c \to 0$$

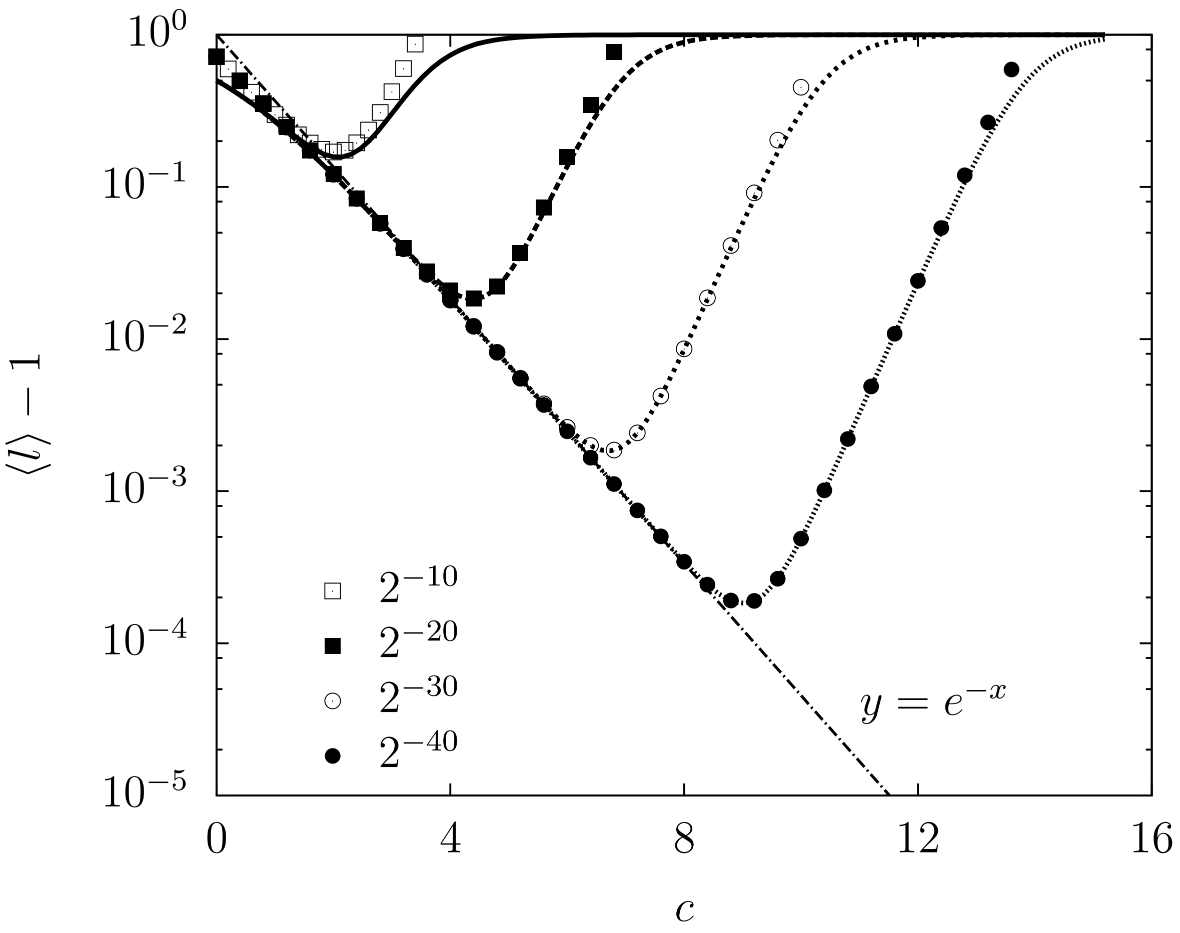
- If the walk starts at distance $d = \alpha L$ from the reference sequence with $\alpha < \frac{1}{2}$, the walk length is nonmonotonic in c and reaches $\ell \to 1$ for $\alpha \to 0$

Mean length of GAW's with Gumbel-distributed randomness



- upper bound: $\ell = 1 + e^c$

Minimum in GAW length with Gumbel-distributed randomness



- $\alpha = 2^{-10} \ldots . 2^{-40}$, location of minimum varies as $c_{\min} \sim \frac{1}{3} \ln(\alpha)$

Summary

- Increasing number of empirical fitness landscapes provide insights into patterns of epistasis

- Existence of accessible pathways is not simply correlated to overall ‘ruggedness’ of the landscape:

  - In the RMF model pathways exist with unity probability for any $c > 0$
  - In the NK-model accessibility vanishes asymptotically for $L \longrightarrow \infty$, possibly at hyperastronomically large values of L

- Static view focused on landscape structure is complemented by dynamic view of accessibility in term of adaptive walks

Summary

- Increasing number of empirical fitness landscapes provide insights into patterns of epistasis

- Existence of accessible pathways is not simply correlated to overall ‘ruggedness’ of the landscape:

  - In the RMF model pathways exist with unity probability for any $c > 0$
  - In the NK-model accessibility vanishes asymptotically for $L \longrightarrow \infty$, possibly at hyperastronomically large values of L

- Static view focused on landscape structure is complemented by dynamic view of accessibility in term of adaptive walks

Thanks to:
Jasper Franke, Johannes Neidhart, Stefan Nowak, Benjamin Schmiegelt,
Ivan Szendro (Cologne)
Arjan de Visser (Wageningen), Su-Chan Park (Seoul)