

Modeling the major transitions in the evolution of language

Willem Zuidema

Language Evolution & Computation Research Unit

Institute for Cell, Animal and Population Biology

University of Edinburgh, U.K.

<http://www.ling.ed.ac.uk/~jelle>

jelle@ling.ed.ac.uk

Foundations of Language

(Ray Jackendoff, 2002, Oxford University Press)

- UG as a toolbox, with many components
- lexicalisation
- redundancy
- cross-linguistic perspective
- developmental scenario
- evolutionary scenario
- processing

phonology	semantics	syntax
$/ f \in R f e /$	$\lambda y \lambda x \text{Search}(x, y)$	$(n \setminus s) / n$

I will not inquire as to the details of how increased expressive power came to spread through a population [...] nor how the genome and the morphogenesis of the brain accomplished these changes. Accepted practice in evolutionary psychology [...] generally finds it convenient to ignore these problems; I see no need at the moment to hold myself to a higher standard than the rest of the field. (Jackendoff, 2002, Foundations of Language, p. 237)

- Problems of *altruism* and *coordination*

Overview

1. The major transitions
2. The Emergence of Compositionality
 - (a) Natural Selection (Nowak & Krakauer, 1999)
 - (b) Iterated Learning (Kirby, 2000; Zuidema, 2003)
 - (c) shortcomings
3. Extended Formalism
4. Results
5. Conclusions

Natural Selection

Formalism of Hurford (1989, *Lingua*), Oliphant (1996, PhD-thesis UCSD)

$$S = \left(\begin{array}{c|ccccc} & \text{sent signal} & & & & \\ \text{intention } \downarrow & 1\text{kHz} & 2\text{kHz} & 3\text{kHz} & 4\text{kHz} & 5\text{kHz} \\ \hline \textit{eagle approaches} & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ \textit{snake approaches} & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \\ \textit{tiger approaches} & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{array} \right)$$

$$R = \left(\begin{array}{c|ccc} & \text{interpretation} & & \\ \text{received signal } \downarrow & \textit{eagle approaches} & \textit{snake approaches} & \textit{tiger approaches} \\ \hline 1\text{kHz} & 1.0 & 0.0 & 0.0 \\ 2\text{kHz} & 1.0 & 0.0 & 0.0 \\ 3\text{kHz} & 0.0 & 1.0 & 0.0 \\ 4\text{kHz} & 0.0 & 0.0 & 1.0 \\ 5\text{kHz} & 0.0 & 0.0 & 1.0 \end{array} \right)$$

$$U = \left(\begin{array}{c|ccccc} & \text{received signal} & & & & \\ \text{sent signal} \downarrow & 1\text{kHz} & 2\text{kHz} & 3\text{kHz} & 4\text{kHz} & 5\text{kHz} \\ \hline 1\text{kHz} & 0.7 & 0.2 & 0.1 & 0.0 & 0.0 \\ 2\text{kHz} & 0.2 & 0.6 & 0.2 & 0.0 & 0.0 \\ 3\text{kHz} & 0.0 & 0.2 & 0.6 & 0.2 & 0.0 \\ 4\text{kHz} & 0.0 & 0.0 & 0.2 & 0.6 & 0.2 \\ 5\text{kHz} & 0.0 & 0.0 & 0.1 & 0.2 & 0.7 \end{array} \right)$$

The fitness is given by (Nowak & Krakauer, 1999):

$$F(L, L') = \frac{1}{2} \sum_{m=1}^M \sum_{f=1}^F \left[S_{mf} \left(\sum_{f'=1}^F U'_{ff'} R'_{f'm} \right) + S'_{mf} \left(\sum_{f'=1}^F U_{ff'} R_{f'm} \right) \right]$$

Assume the world consists of objects and actions, a fraction ϕ of which is relevant. Nowak & Krakauer (1999) show that with sufficiently high ϕ the maximum fitness of compositional languages is higher. I.e.

$$F(L^+, L^+) > F(L^-, L^-).$$

- However, crucial is that there is a path of *ever increasing fitness* from non-compositional (L^-) to compositional (L^+) languages. I.e.

$$\begin{aligned} F(L^+, L^+) &> F(L^+, L^-) > F(L^-, L^-) \\ F(L^+, L^+) &> \dots > F(L^-, L^-). \end{aligned}$$

Mixed Strategies

$$S = \left(\begin{array}{c|cccccc} & \text{sent signal} & & & & & \\ \text{intention } \downarrow & A & B & C & ab & cb & ad \\ \hline 1 \text{ eagle approaches} & 1-x & 0.0 & 0.0 & x & 0.0 & 0.0 \\ 2 \text{ snake approaches} & 0.0 & 1-x & 0.0 & 0.0 & x & 0.0 \\ 3 \text{ eagle leaves} & 0.0 & 0.0 & 1-x & 0.0 & 0.0 & x \end{array} \right)$$

$$R = \left(\begin{array}{c|ccc} \text{sent } \downarrow & 1 & 2 & 3 \\ \hline A & 1 & 0 & 0 \\ B & 0 & 1 & 0 \\ C & 0 & 0 & 1 \\ ab & 1 & 0 & 0 \\ cb & 0 & 1 & 0 \\ ad & 0 & 0 & 1 \end{array} \right)$$

$$U = \left(\begin{array}{c|cccccc} & \text{received signal} & & & & & \\ \text{sent } \downarrow & A & B & C & ab & cb & ad \\ \hline A & . & * & * & 0 & 0 & 0 \\ B & * & . & * & 0 & 0 & 0 \\ C & * & * & . & 0 & 0 & 0 \\ ab & 0 & 0 & 0 & * & . & . \\ cb & 0 & 0 & 0 & . & * & . \\ ad & 0 & 0 & 0 & . & . & * \end{array} \right)$$

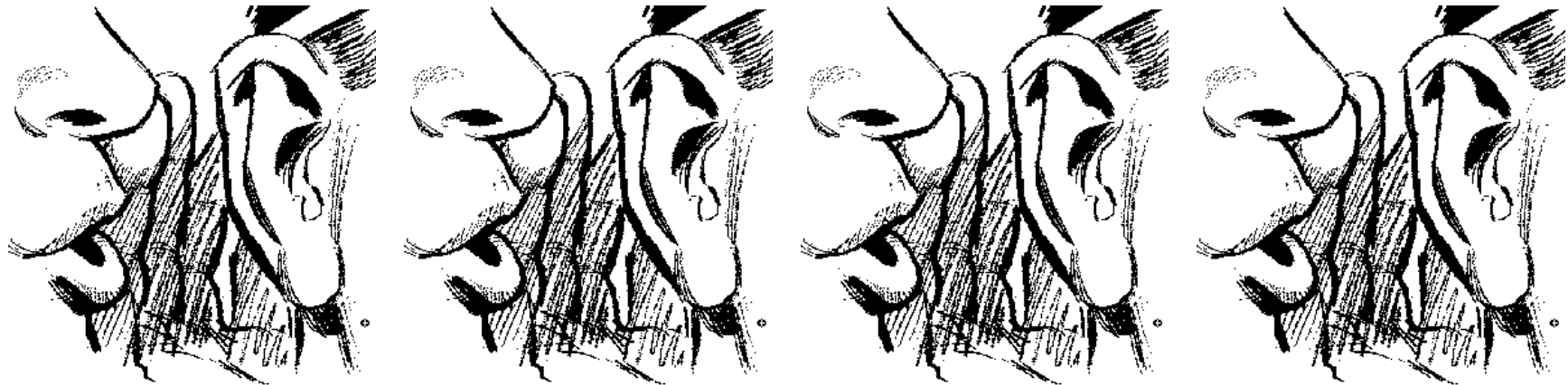
N & K show that for mixed strategies, more compositional languages will always do better:

$$F(L', L') > F(L', L) > F(L, L) \text{ if } x' > x > 0.$$

- However, cost of additional system (memory, confusion) and temporal dimension of holistic signals are completely ignored.

Iterated Learning

Languages are transmitted culturally. The elements of language compete with each other for survival. Compositional rules will outcompete non-compositional ones, because they can be inferred from incomplete data. (Hare & Elman, 1995; Kirby, 2000; Brighton, 2003).



Language Representation

(e.g. Zuidema, 2003, NIPS Proceedings)

Context-free grammars of the form: $A \mapsto t$, $A \mapsto BC$ or $A \mapsto Bt$. Convenient, and no restrictions on expressiveness.

Parsing: depth-first search, with maximum depth d

Derivation: random string from parsable language

Grammar Induction

Incorporation: extend the language, such that it includes the encountered string

Compression: substitute frequent and long substrings with a nonterminal (the grammar becomes smaller and the language remains unchanged)

Generalization: equate two nonterminals if they occur frequently in the same context (sufficient expressiveness enforced)

Generation 1

S \mapsto h0033 (1) S \mapsto g2110 (6) S \mapsto 20113 (1) S \mapsto g112 (6) S \mapsto gg (36) S \mapsto g0131
(6) S \mapsto g0303 (6) S \mapsto g123 (6) S \mapsto gf (0) S \mapsto 210012 (1) S \mapsto 032133 (1)
S \mapsto 312031 (1) S \mapsto 113201 (1) S \mapsto 021132 (1) g \mapsto 30 (17) g \mapsto f (17) g \mapsto h2
(17) g \mapsto 311 (17) g \mapsto 322 (17) f \mapsto 222 (17) g \mapsto 033 (17) h \mapsto 00 (18)

Generation 20

S \mapsto f002 (5) S \mapsto ff (20) S \mapsto fg (5) S \mapsto fh (5) S \mapsto f033 (5) f \mapsto 2110 (13) f \mapsto 322
(13) f \mapsto 002 (8) f \mapsto 112 (13) f \mapsto 0131 (13) g \mapsto 311 (5) h \mapsto 123 (5)

The emergence of compositionality explained?

Compositionality: the property that the meaning of the whole (e.g. a sentence) is a function of the meaning of the parts (e.g. the words) and the way they are put together.

Existing models require a structured language to be already present in the population before the linguistic innovations can successfully spread in a population.

Topology preservation

I explore a possible route for a structured language to emerge without the capacity for compositionality present in the population. The structure in this case is *topology preservation* between meaning-space and signal-space, i.e. similar meanings are expressed with similar forms (signals).

Can topology preservation emerge as a side-effect of optimising communication under noisy conditions?

A formalism for communication under noisy conditions

- Assume that there are M different meanings that an individual might want to express, and F different signals (forms) that it can use for this task.
- The communication system of an individual is represented with a *production matrix* S (S gives for every meaning m and every signal f , the probability that the individual chooses f to convey m);
- and an *interpretation matrix* R . (R gives for every signal f and meaning m , the probability that f will be interpreted as m).

- Signals can be more or less similar to each other and there is noise on the transmission of signals which depends on these similarities (*confusion matrix U*).
- Meanings can be more or less similar to each other, and the value of a certain *interpretation* depends on how close it is to the *intention* (*value matrix V*)

Example: Vervet monkey alarm calls

Three different types of predators: from the air (eagles), from the ground (tigers) and from the trees (snakes).

The monkeys are capable of making a number (say 5) of different sounds that range on one axis (e.g. pitch, from high to low) and are more easily confused if they are closer together.

If one makes a mistake, typically not every mistake is equally bad.

$$U = \left(\begin{array}{c|ccccc} & \text{received signal} & & & & \\ \text{sent signal} \downarrow & 1\text{kHz} & 2\text{kHz} & 3\text{kHz} & 4\text{kHz} & 5\text{kHz} \\ \hline 1\text{kHz} & 0.7 & 0.2 & 0.1 & 0.0 & 0.0 \\ 2\text{kHz} & 0.2 & 0.6 & 0.2 & 0.0 & 0.0 \\ 3\text{kHz} & 0.0 & 0.2 & 0.6 & 0.2 & 0.0 \\ 4\text{kHz} & 0.0 & 0.0 & 0.2 & 0.6 & 0.2 \\ 5\text{kHz} & 0.0 & 0.0 & 0.1 & 0.2 & 0.7 \end{array} \right)$$

$$V = \left(\begin{array}{c|ccc} & \text{intentions} & & \\ \text{interpretations} \downarrow & \textit{eagle} & \textit{snake} & \textit{tiger} \\ \hline \textit{eagle} & 0.9 & 0.5 & 0.1 \\ \textit{snake} & 0.2 & 0.9 & 0.2 \\ \textit{tiger} & 0.1 & 0.5 & 0.9 \end{array} \right)$$

$$S = \left(\begin{array}{c|ccccc} & \text{sent signal} & & & & \\ \text{intention} \downarrow & 1\text{kHz} & 2\text{kHz} & 3\text{kHz} & 4\text{kHz} & 5\text{kHz} \\ \hline \textit{eagle} & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ \textit{snake} & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \\ \textit{tiger} & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{array} \right)$$

$$R = \left(\begin{array}{c|ccc} & \text{interpretation} & & \\ \text{received signal} \downarrow & \textit{eagle} & \textit{snake} & \textit{tiger} \\ \hline 1\text{kHz} & 1.0 & 0.0 & 0.0 \\ 2\text{kHz} & 1.0 & 0.0 & 0.0 \\ 3\text{kHz} & 0.0 & 1.0 & 0.0 \\ 4\text{kHz} & 0.0 & 0.0 & 1.0 \\ 5\text{kHz} & 0.0 & 0.0 & 1.0 \end{array} \right)$$

$$F_{ij} = V \cdot \left(S^i \times \left(U \times R^j \right) \right) \quad (1)$$

In this formula, “ \times ” represents the usual matrix multiplication and “ \cdot ” represents dot-multiplication (the sum of all multiplications of corresponding elements in both matrices; the result of dot-multiplication is not a matrix, but a scalar).

$$\begin{aligned} F_{ij} = & 0.7 \times 0.9 + 0.2 \times 0.5 + 0.2 \times 0.5 + 0.6 \times 0.9 \\ & + 0.2 \times 0.5 + 0.1 \times 0.5 + 0.2 \times 0.9 + 0.7 \times 0.9 = 2.33 \end{aligned}$$

Distributed hill-climbing

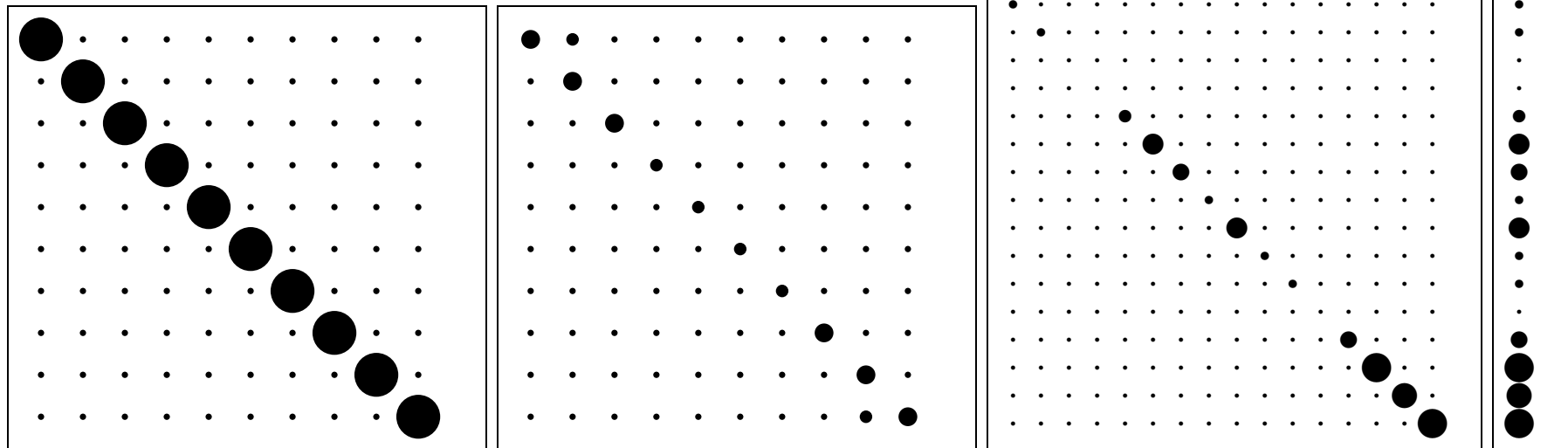
- The values in the S and R matrices are all either 1 or 0
- Distributed hill-climbing:
 1. Random speaker (i) and hearer (j) are picked, and F_{ij} is measured;
 2. A random change is made in a random matrix of the speaker (or hearer), and F_{ij} is measured again;
 3. If the F_{ij} is better, the change is kept; otherwise, it is reverted.

Motivation

for this style of optimization:

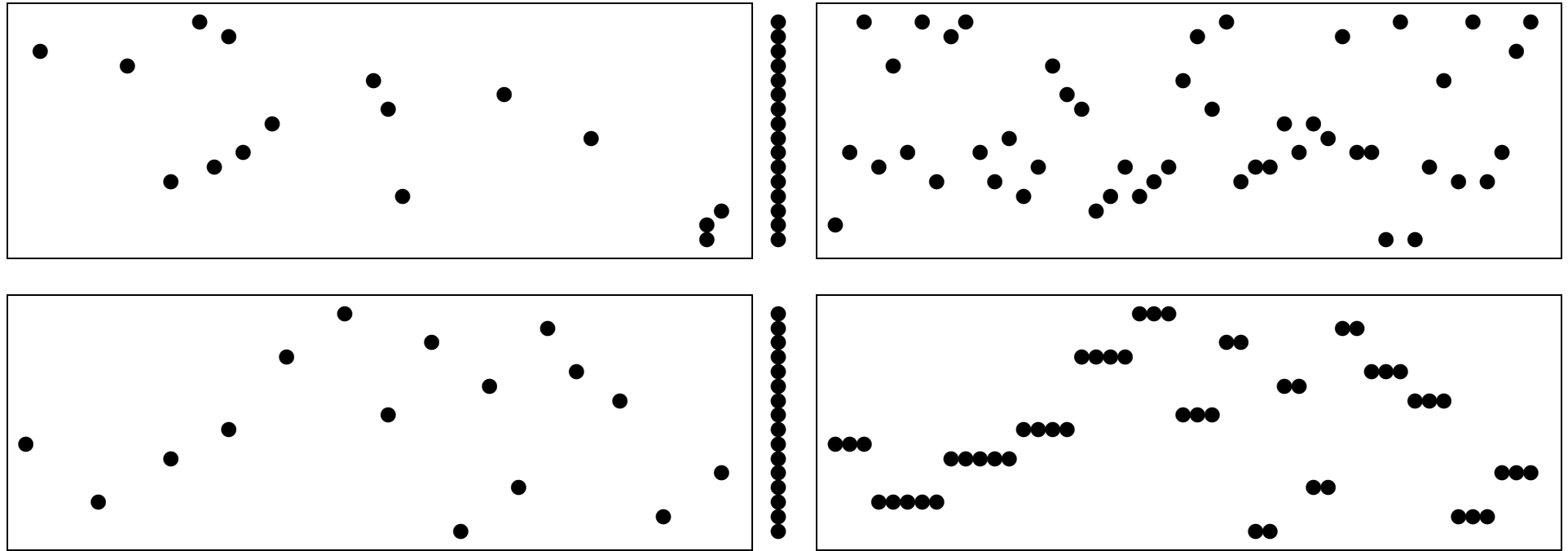
1. it is fast and straightforward to implement;
2. it works well, and gives, if not the optimum, a good insight on characteristics of the optimal communication system;
3. it shows possible *routes* to (near-) optimal communication systems, and in a sense forms an abstraction for both learning and evolution.

Control parameters: V and U -matrices



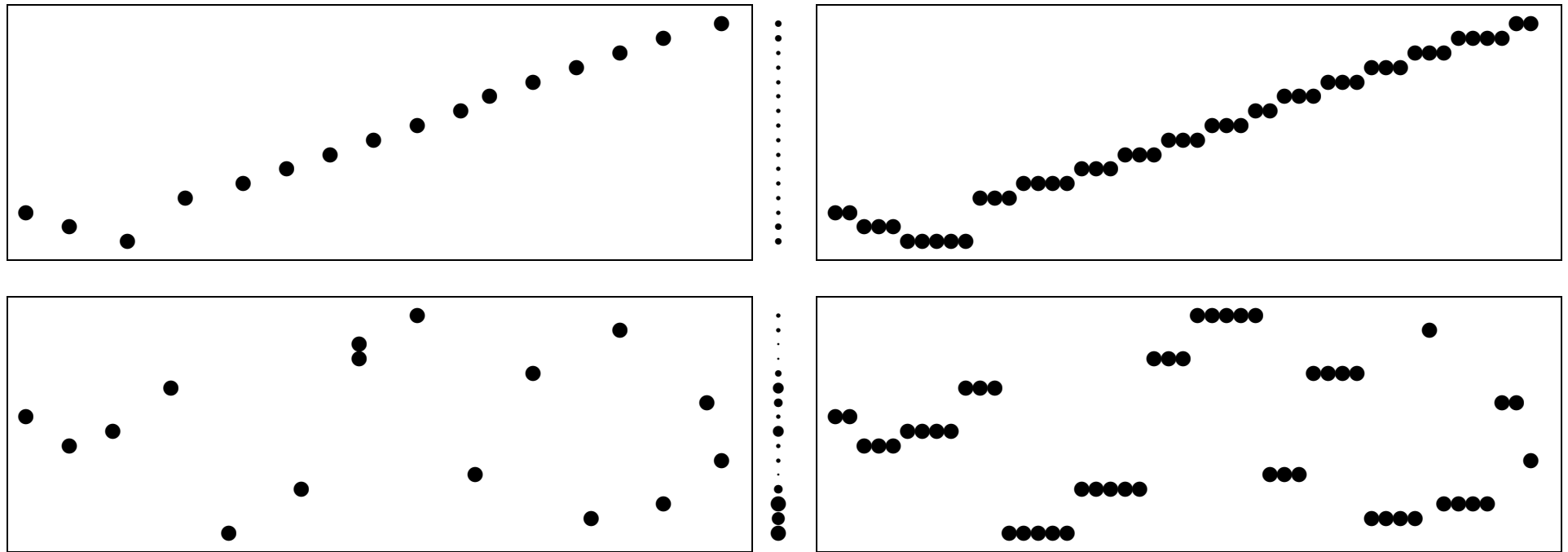
V : 0d, 1d homogeneous, 0d heterogeneous

Results
Specificity, Sharedness, Distinctiveness

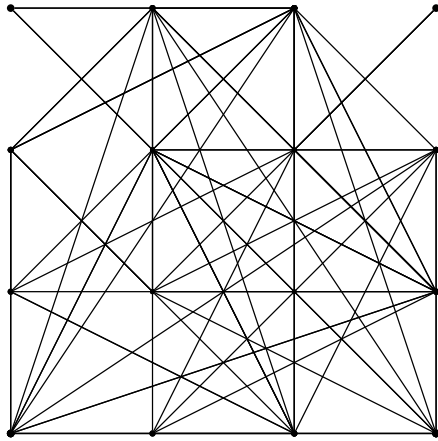


U:1d, V:0d homogeneous, $t = 0$, $t = \infty$

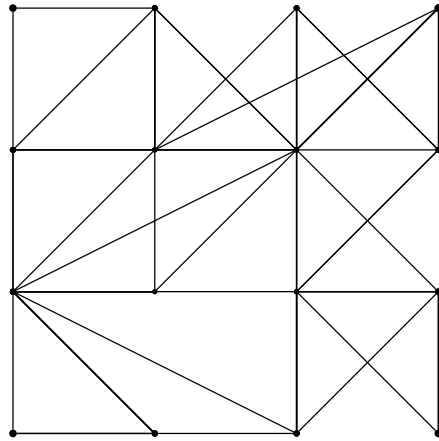
Topology preservation, Heterogeneity



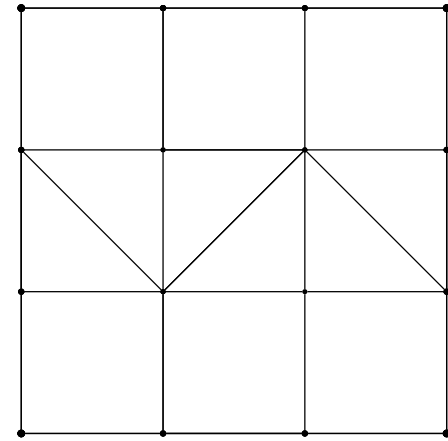
U:1d, V:0d homogeneous/heterogeneous, $t = \infty$



(a) $t=0$

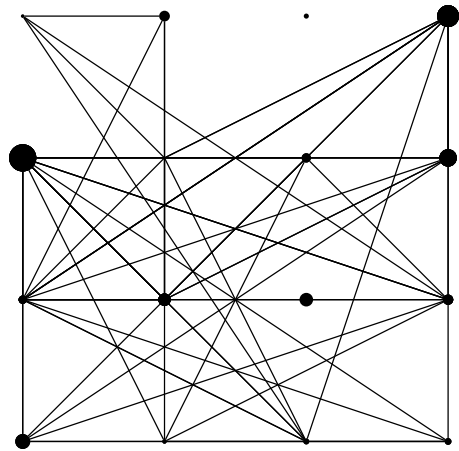


(b) $t=10^6$

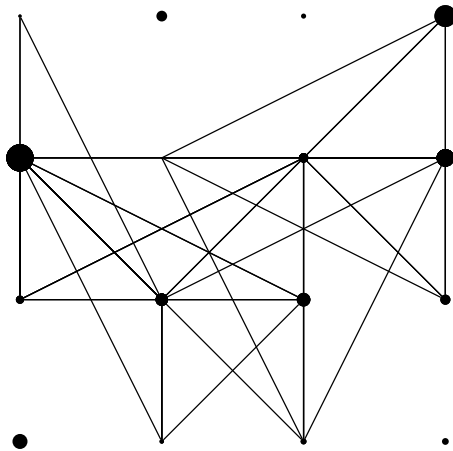


(c) $t=\infty$

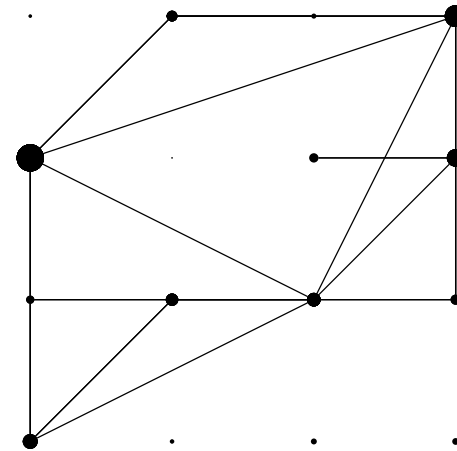
U:2d, V:2d homogeneous, $t = 0, 10^6, \infty$



(a) $t=0$

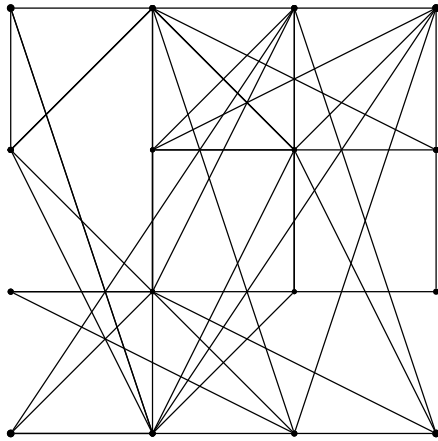


(b) $t=10^6$

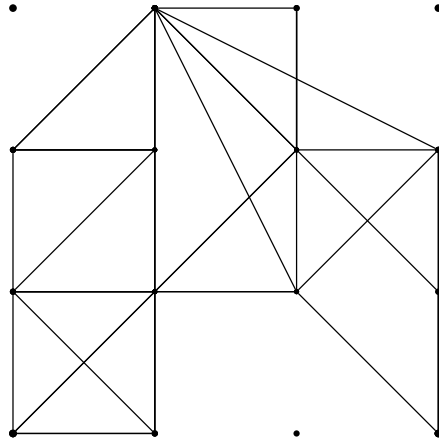


(c) $t=\infty$

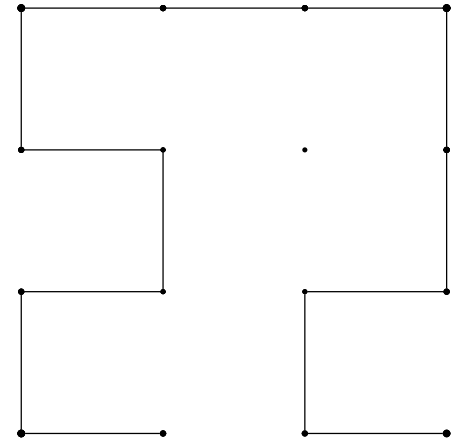
U:2d, V:2d heterogeneous, $t = 0, 10^6, \infty$



(a) $t=0$

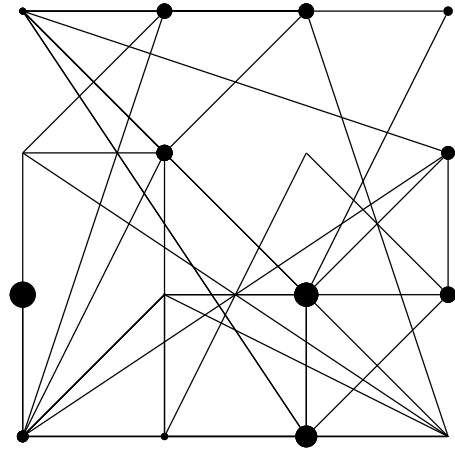


(b) $t=10^6$

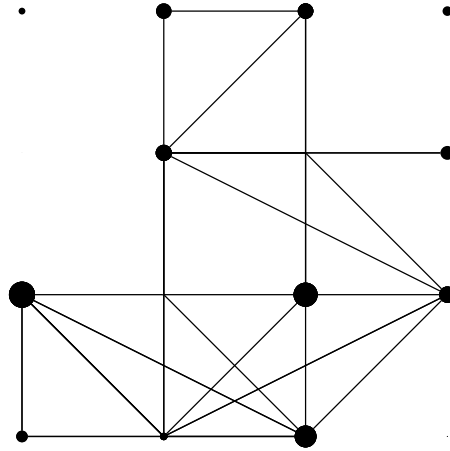


(c) $t=\infty$

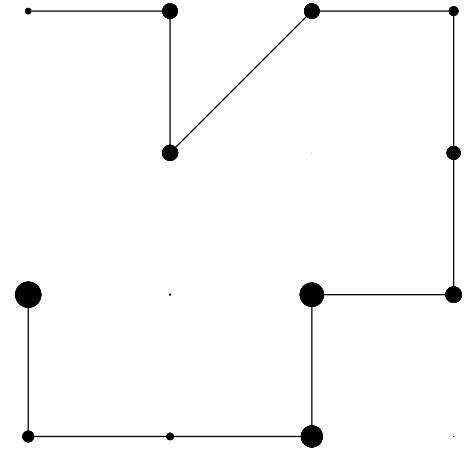
U:1d, V:2d homogeneous, $t = 0, 10^6, \infty$



(a) $t=0$



(b) $t=10^6$



(c) $t=\infty$

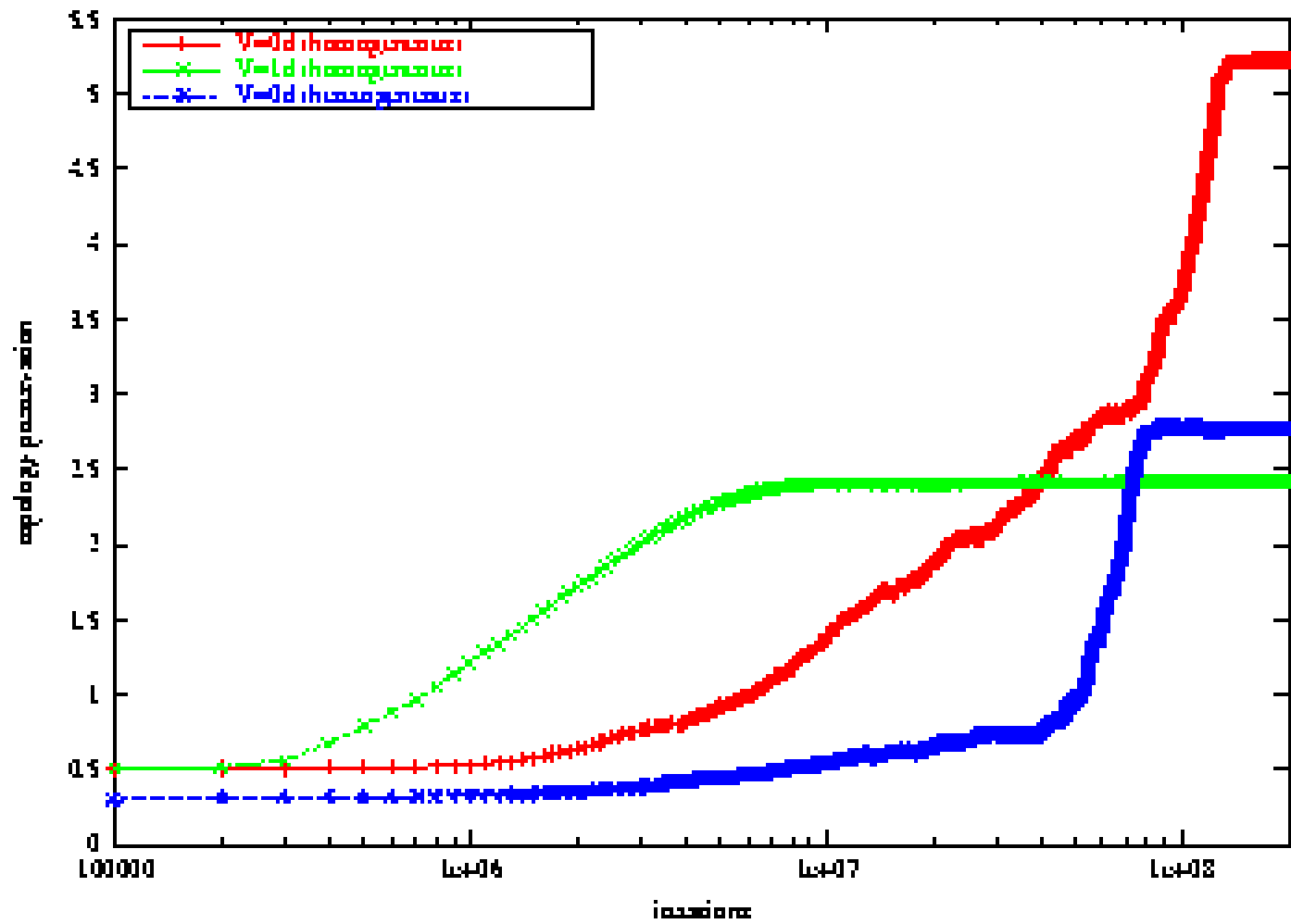
U:1d, V:2d heterogeneous, $t = 0, 10^6, \infty$

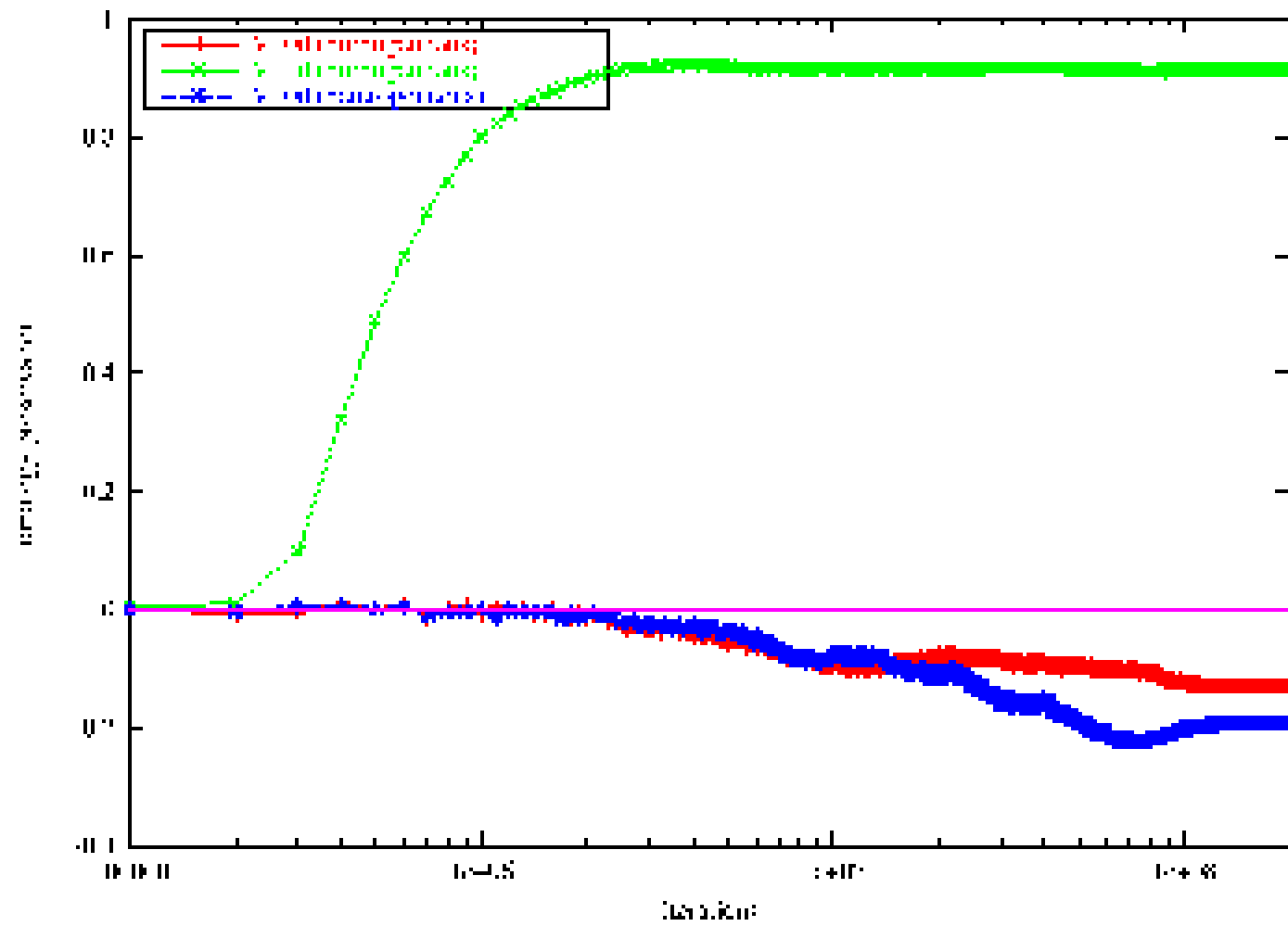
Summary of results

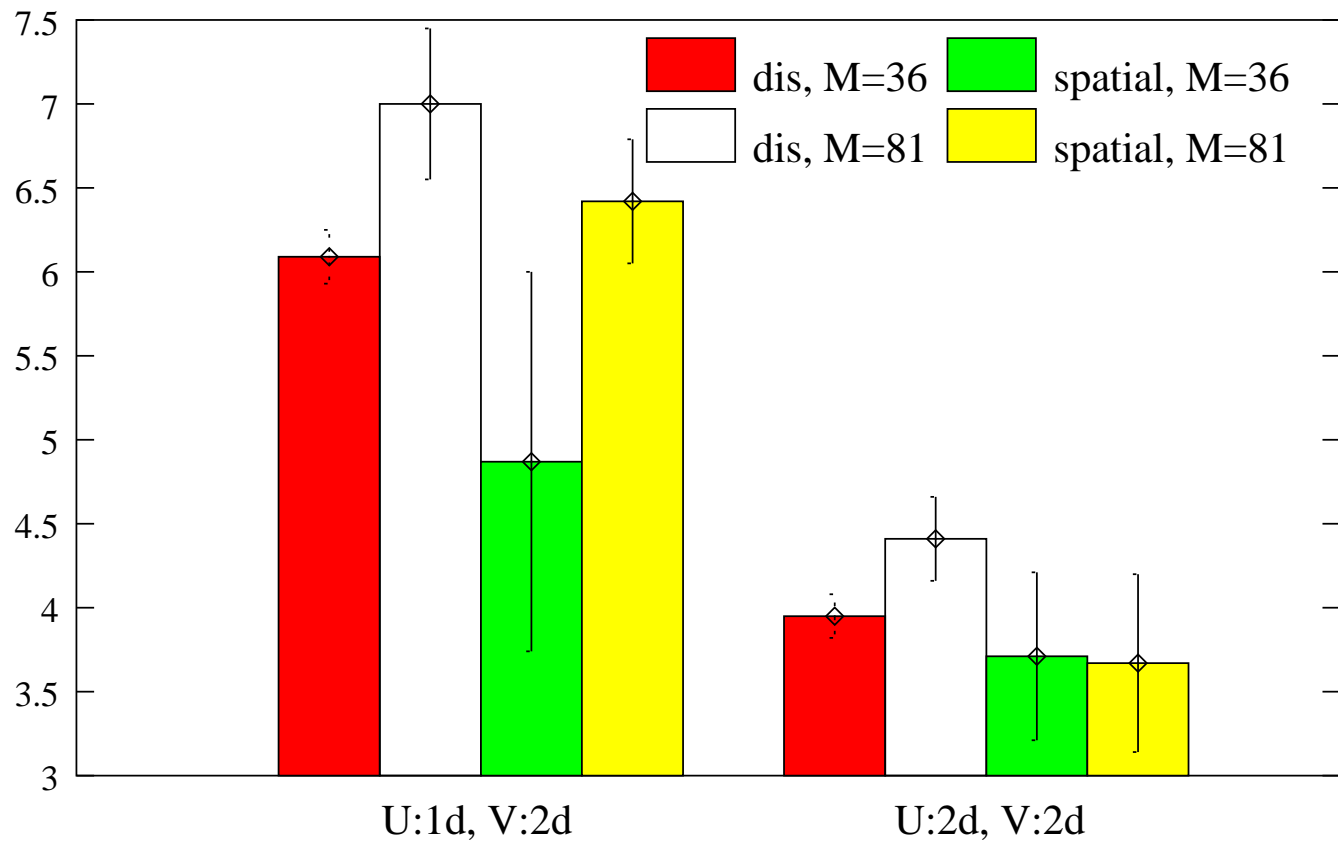
- Specificity
- Sharedness
- Distinctiveness
- Topology preservation
- High-value meanings first
- Low-value meanings sacrificed

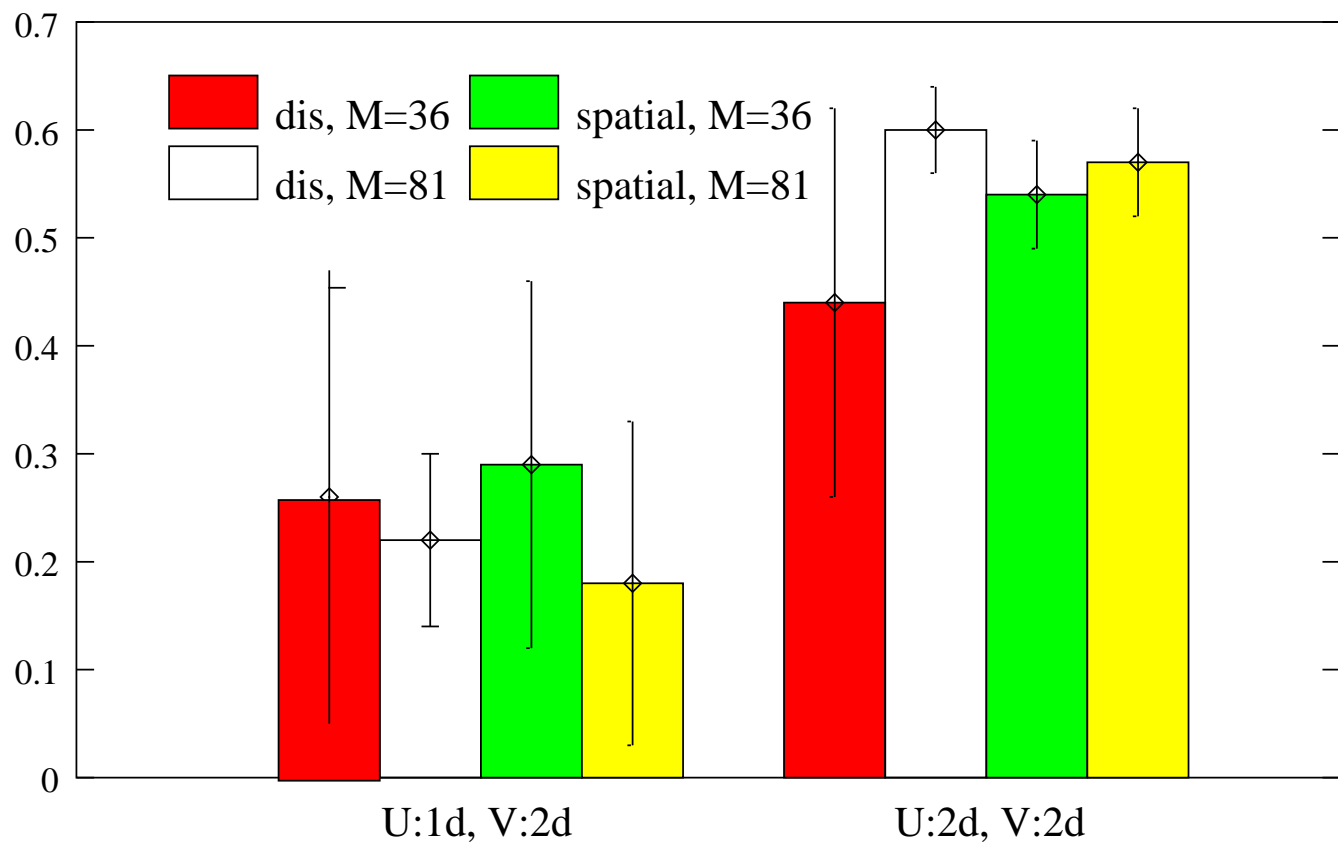
Measuring topology preservation

$$C = \sum_{m, m' \in M} \text{correlation} (D (m, m'), D (S[m], S[m'])), \quad (2)$$









Discussion

- Evolutionary stable strategies
- Information theory / channel capacity
- Evolution of compositionality
- Sound symbolism?

Conclusions

1. Crucial for evolutionary explanations of all aspects of language, is to explain how linguistic innovations can spread in a population; showing a better end result is neither sufficient nor necessary;
2. Including plausible assumptions on noise in signalling and a topology in the meaning space, as studied in a simple simulation, opens up the possibility for rich pattern formation that was overlooked in previous robotic and mathematical models;

3. Combinatorial patterning as a strategy to minimize the effects of noise is a possible precursor for *productive* combination;
4. A rich formalism allows for side-effects in evolutionary optimization; side-effects of one adaptation (e.g. learning, noise robustness) might facilitate the next (e.g. phonemic coding, compositionality).

Acknowledgments

- Simon Kirby, Jim Hurford (LEC, Edinburgh), Nick Barton (ICAPB, Edinburgh)
- Gert Westermann, Oxford Brooks (\mapsto *Artificial Life*)
- Bart de Boer, VUB Brussels (\mapsto *BBS*)