icelab | INTEGRATED SCIENCE LAB

UMEÅ UNIVERSITY

Mapping network flows

with incomplete information

Martin Rosvall


Daniel Edler


Jelena Smiljanić


Christopher Blöcker

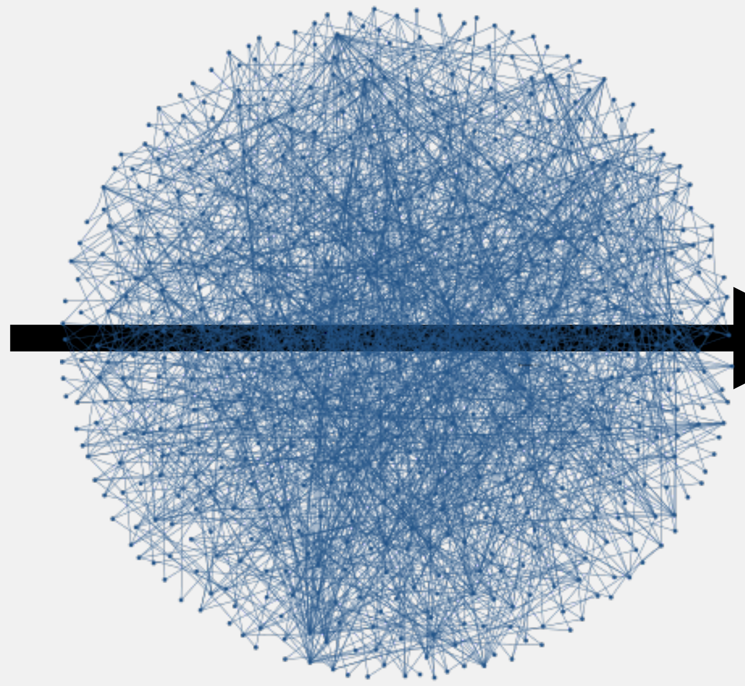

Anton Eriksson

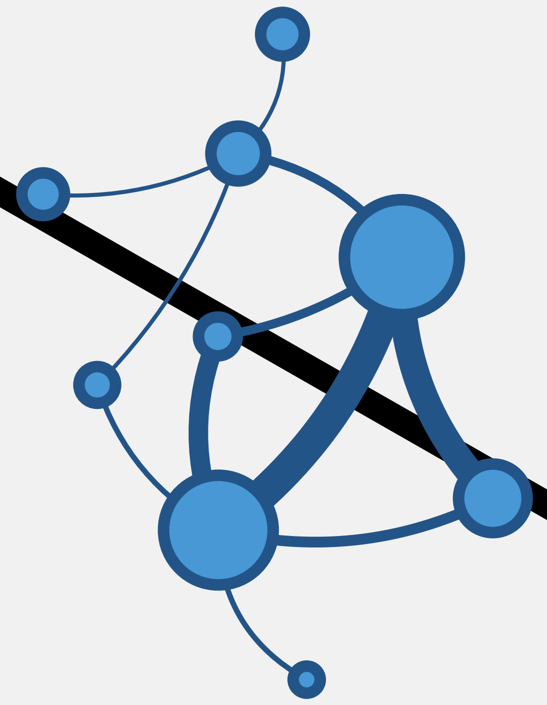

Martin Rosvall

Research cycles in our interdisciplinary group

Research questions

How can we explain natural phenomena X?





How can we explain natural phenomena Y?

Network methods

Method A

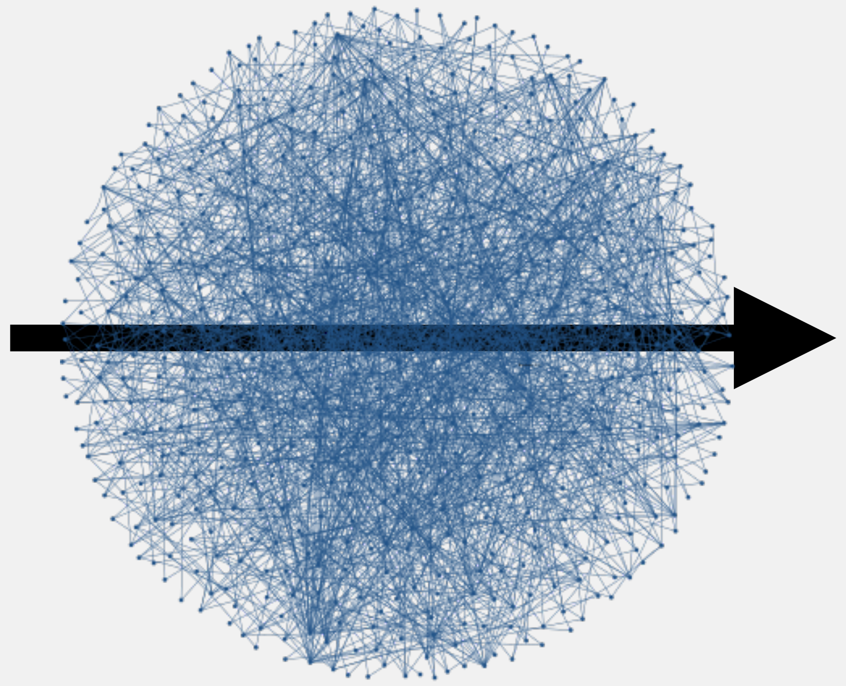
Method B

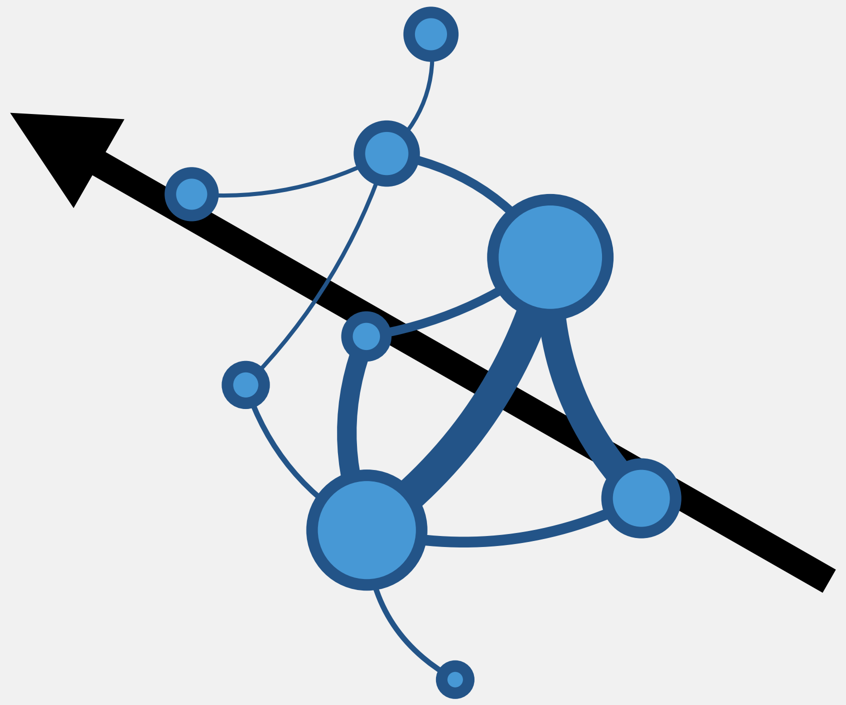
Research cycles in our interdisciplinary group

Research questions

How can we explain natural phenomena Y?

How can we explain natural phenomena Z?



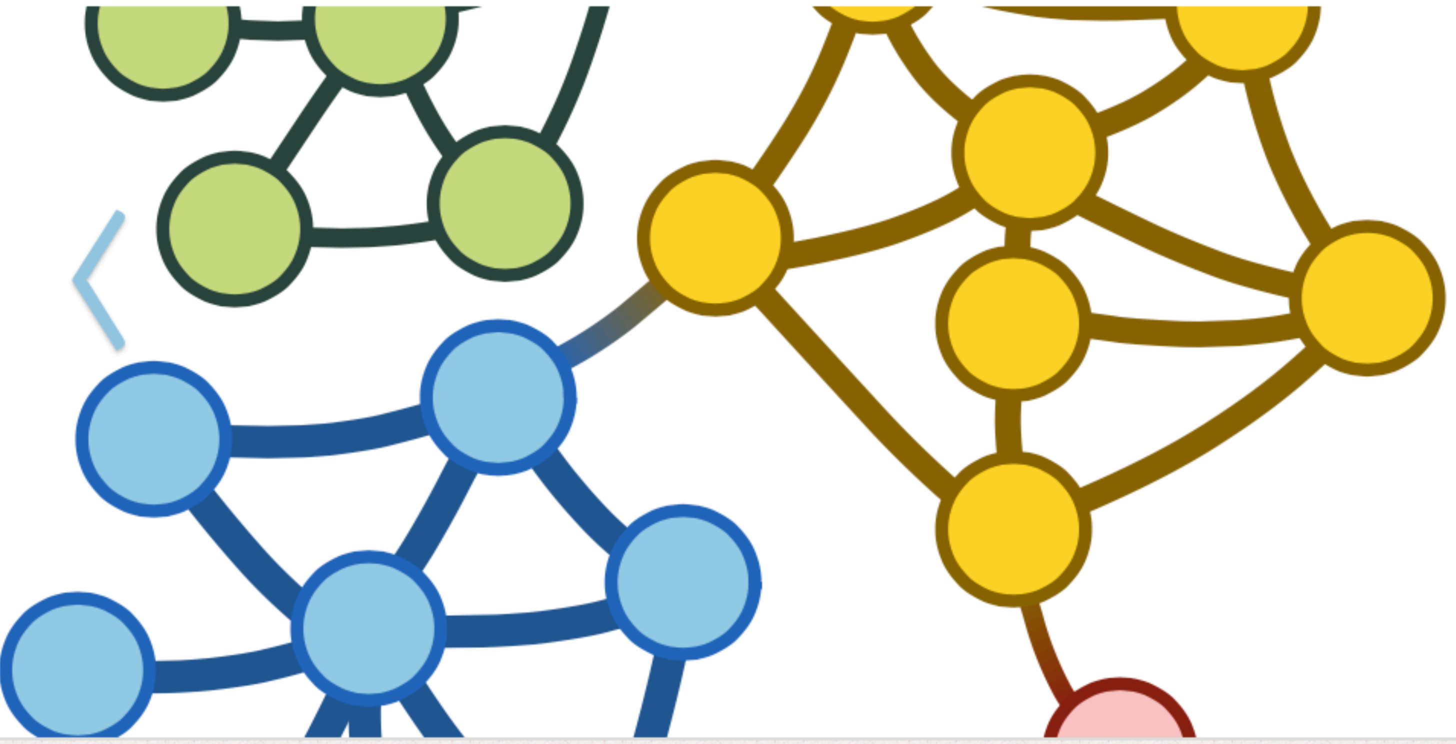


Network methods

Method B

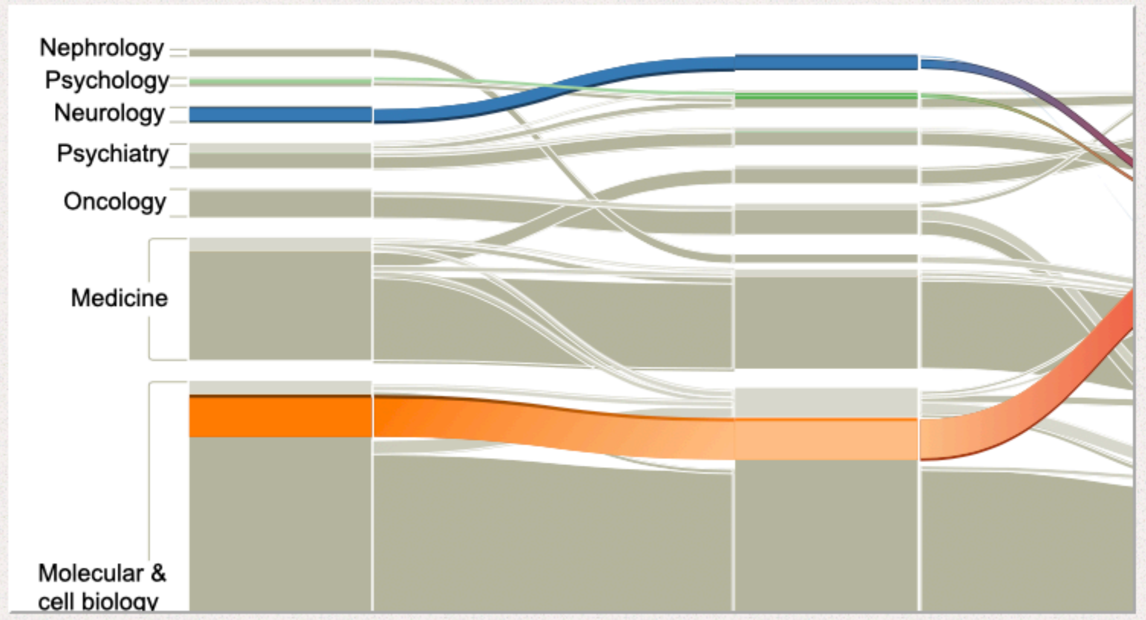
Method C

Explore the mechanics of the map equation



$$L(\mathsf{M}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^{i}_{\circlearrowright} H(\mathcal{P}^{i})$$

Apps »



Code »

```
from infomap import Infomap
im = Infomap()
im.read_file("ninetriangles.net")
im.add_link(1, 10)
im.run("--two-level --num-trials 5")
print(im.codelength)
for node in im.tree:
  if node.is_leaf:
    print(node.node_id, node.module_id)
```

Publications »

Maps of information flow reveal community structure in complex networks

Martin Rosvall and Carl T. Bergstrom

PNAS **105**, 1118 (2008). [arXiv:0707.0609]

To comprehend the multipartite organization of large-scale biological and social systems, we introduce a new information-theoretic approach to reveal community structure in

News

Oct 14, 2021 Release — Infomap binaries — Infomap binaries are now available for Windows, MacOS, and Linux. We also build binary wheels for Windows and macOS.

Oct 4, 2021 Release — Infomap v1.7 — Updated Python API, documentation, and bug fixes (changelog)

Sep 22, 2021 Release — Infomap v1.5 — Updated Python API, bug fixes, CSV and JSON output (changelog)

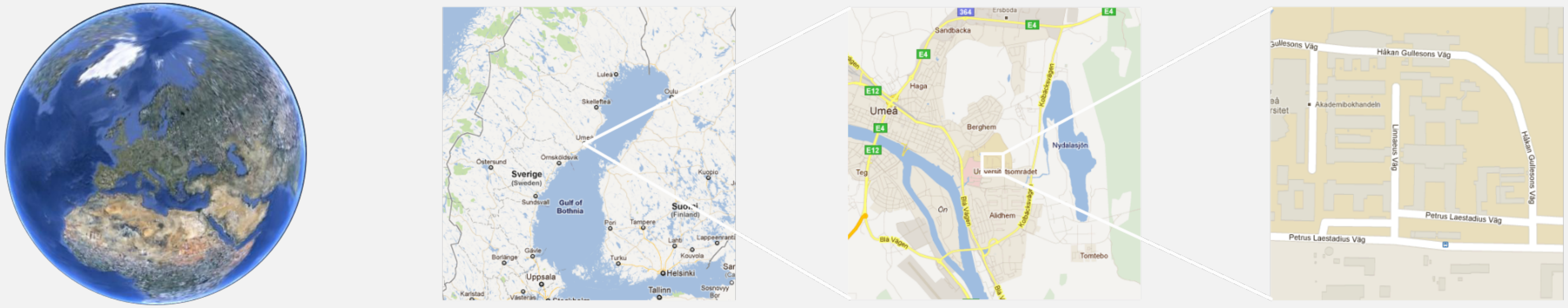
Jun 11, 2021 Research Paper — How choosing random-walk model and network representation matters for flow-based community detection in hypergraphs — Comm. Phys. 4, 133 (2021)

May 11, 2021 Preprint — Flow-based community detection in hypergraphs — arXiv:2105.04389

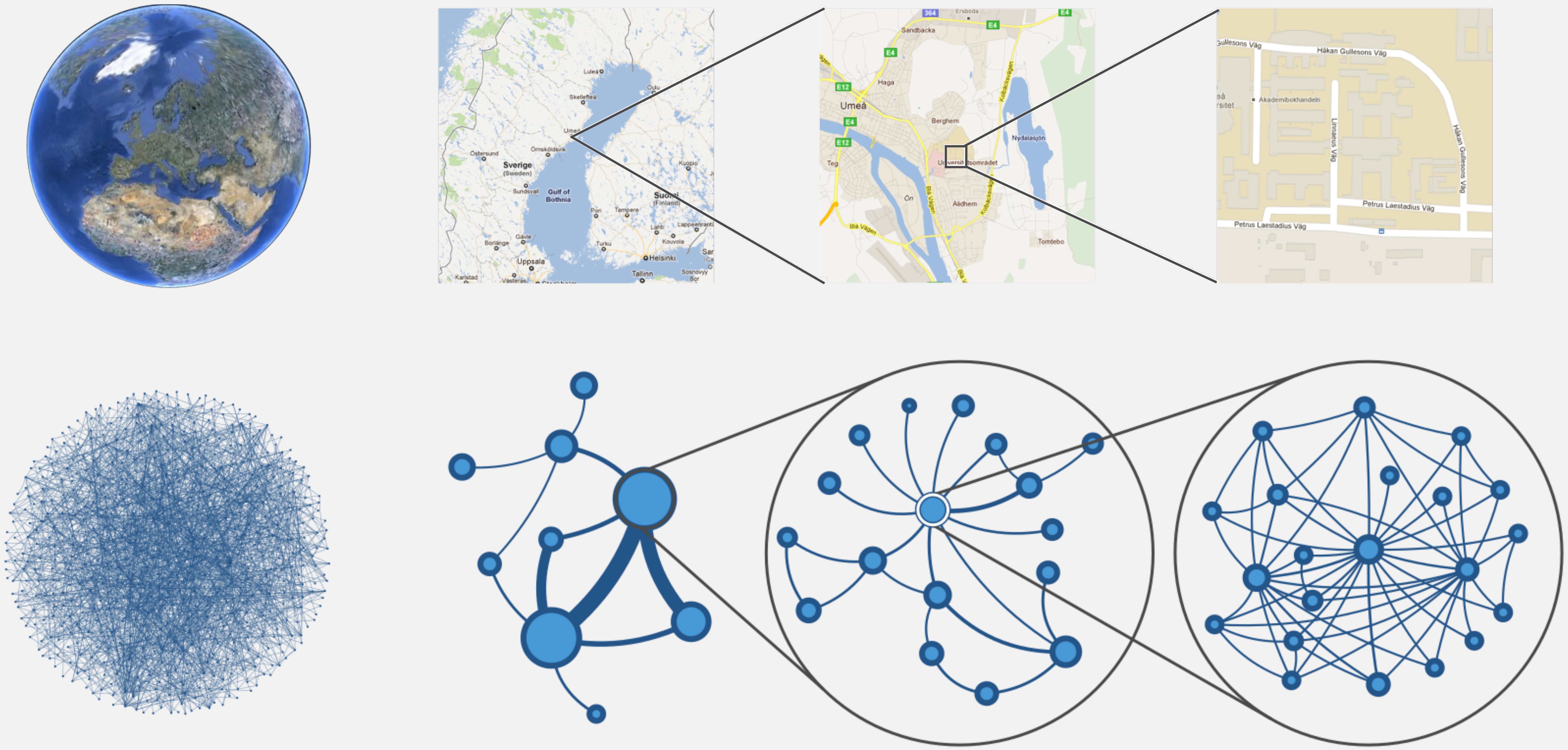
Nov 11, 2020 Research paper — Mapping flows on bipartite networks — Phys. Rev. E 102, 052305 (2020)

Sep 16, 2020 Release — Infomap on Docker Hub — Run Infomap on any operating system with Docker

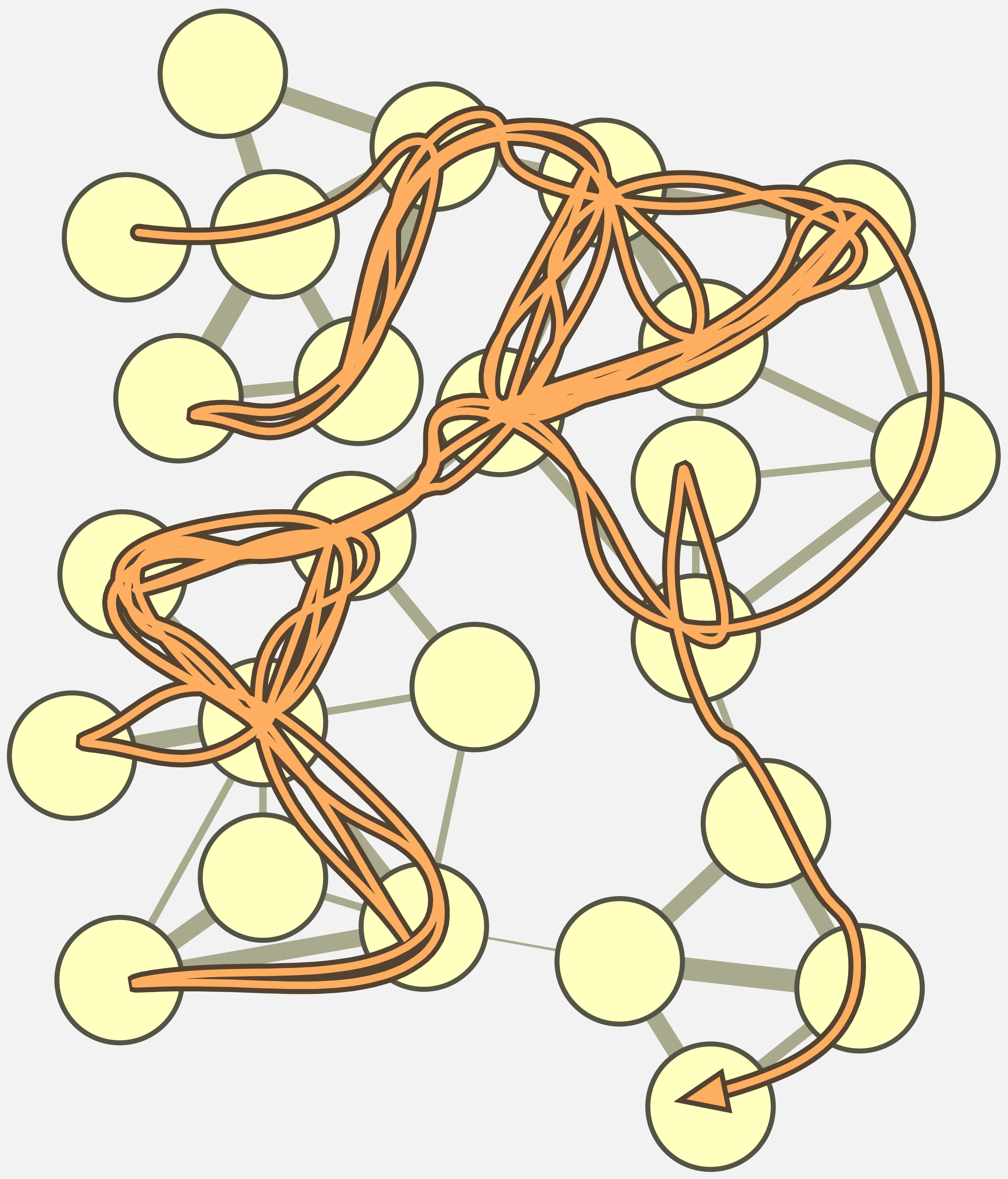
Mapping network flows

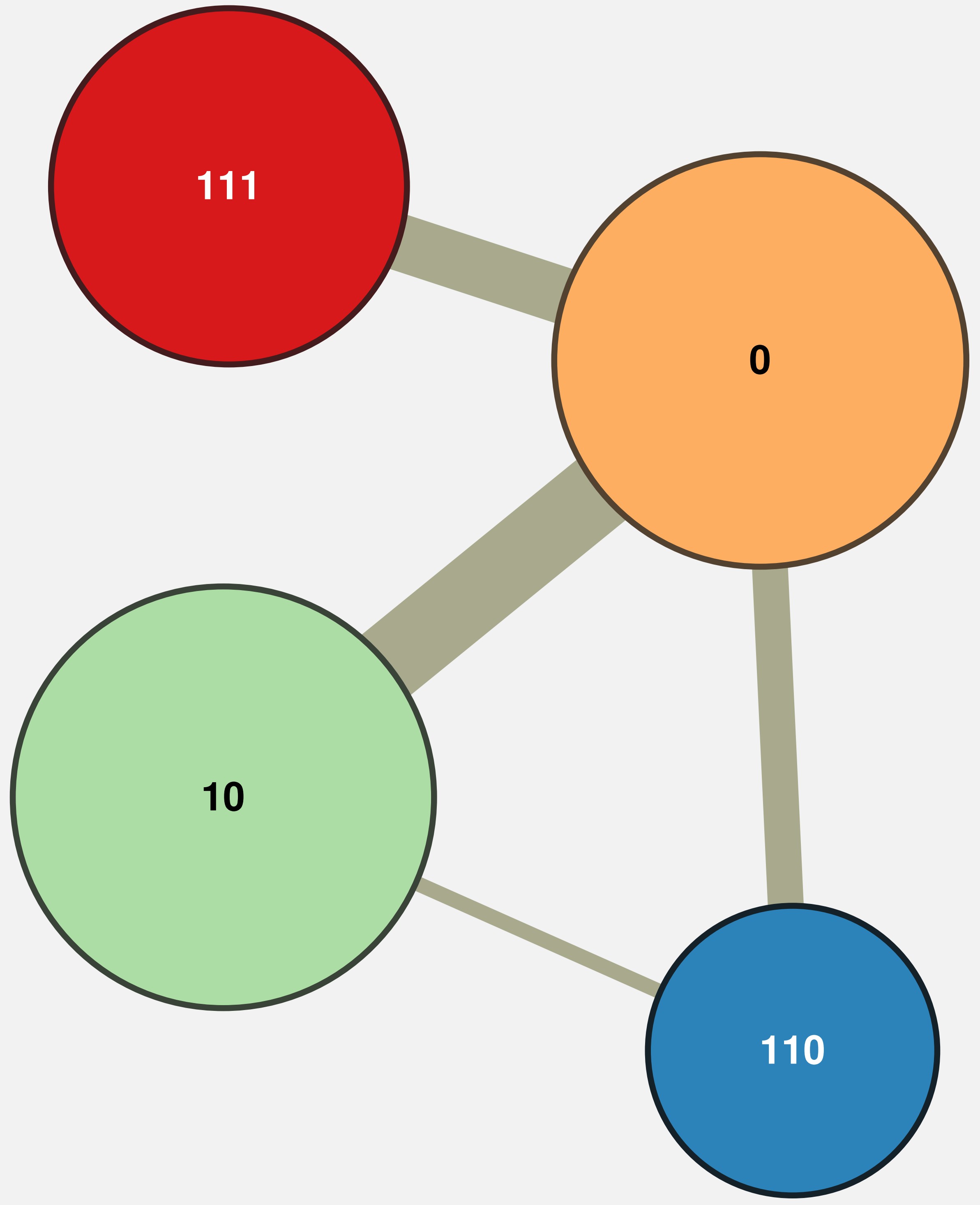


Mapping network flows



Mapping network flows





1. Coding theory

2. Mapping network flows

3. …with incomplete information

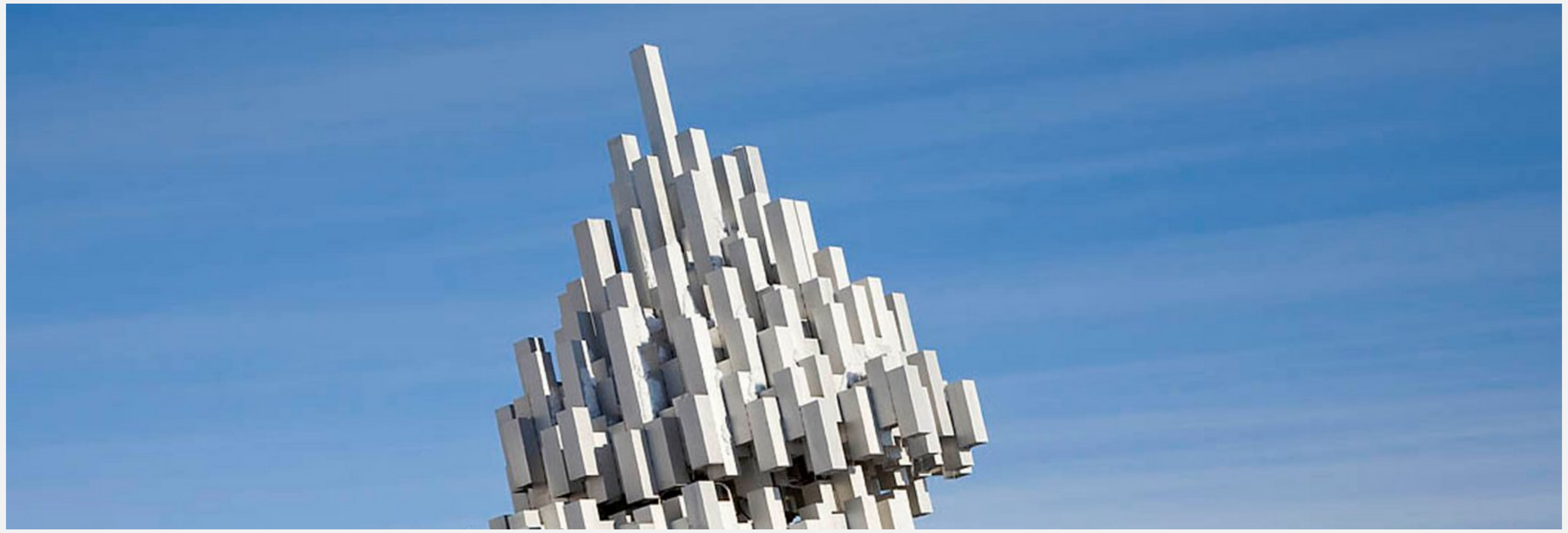
1. **Coding theory**: The minimum description length principle

International Morse Code

1. The length of a dot is one unit.
2. A dash is three units.
3. The space between parts of the same letter is one unit.
4. The space between letters is three units.
5. The space between words is seven units.

A	● ▬	U	● ● ▬
B	▬ ● ● ●	V	● ● ● ▬
C	▬ ● ▬ ●	W	● ▬ ▬
D	▬ ● ●	X	▬ ● ● ▬
E	●	Y	▬ ● ▬ ▬
F	● ● ▬ ●	Z	▬ ▬ ● ●
G	▬ ▬ ●		
H	● ● ● ●		
I	● ●		
J	● ▬ ▬ ▬		
K	▬ ● ▬	1	● ▬ ▬ ▬ ▬
L	● ▬ ● ●	2	● ● ▬ ▬ ▬
M	▬ ▬	3	● ● ● ▬ ▬
N	▬ ●	4	● ● ● ● ▬
O	▬ ▬ ▬	5	● ● ● ● ●
P	● ▬ ▬ ●	6	▬ ● ● ● ●
Q	▬ ▬ ● ▬	7	▬ ▬ ● ● ●
R	● ▬ ●	8	▬ ▬ ▬ ● ●
S	● ● ●	9	▬ ▬ ▬ ▬ ●
T	▬	0	▬ ▬ ▬ ▬ ▬

5.8MB (tiff) → 0.91MB (tiff + LZW)



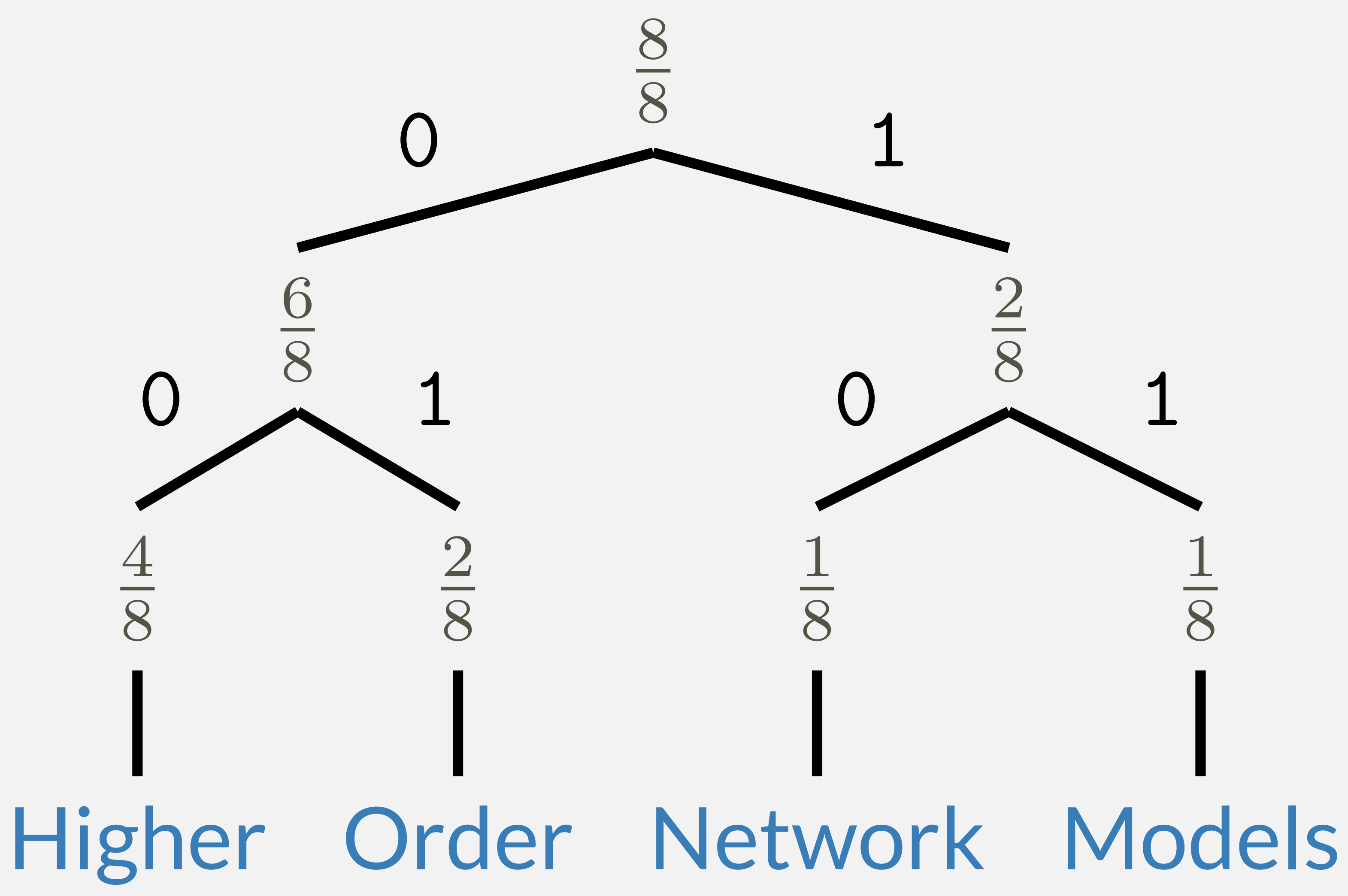
5.8MB (tiff) → 2.8MB (tiff + LZW)

$X = \{\text{Higher}, \text{Order}, \text{Network}, \text{Models}\}$
$\mathcal{P} = \{P(\text{Higher}) = \frac{1}{2}, P(\text{Order}) = \frac{1}{4}, P(\text{Network}) = \frac{1}{8}, P(\text{Models}) = \frac{1}{8}\}$

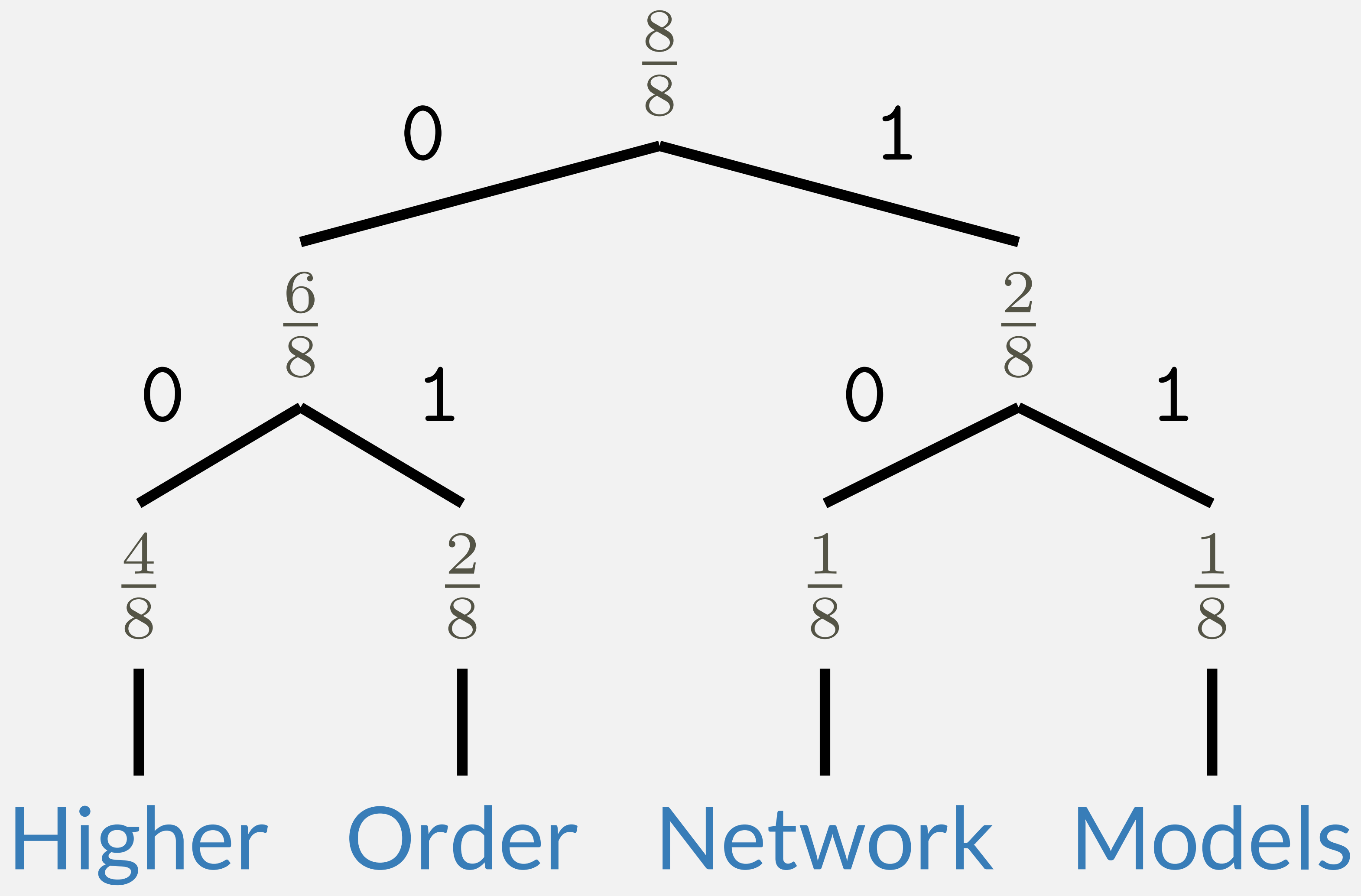
Higher Network Network Higher Higher Higher Order Network Higher Higher Network Higher Network Higher Higher Order Higher Models Models Higher Higher Order Models Higher Higher Higher Higher Higher Higher Higher Higher Higher Higher Higher Higher Higher Network Order Order Higher Models Network Higher Higher Order Order Models Higher Network Order Higher Order Models Network Order Order Higher Higher Network Higher Higher Order Order Order Higher Higher Order Network Order Higher Higher Higher Higher Order Models Order Higher Higher Order Models Network Order Network Higher Order Models Order Higher Higher Order Models Higher Network Network Models Order Network Order Order

Higher = 00, Order = 01, Network = 10, Models = 11

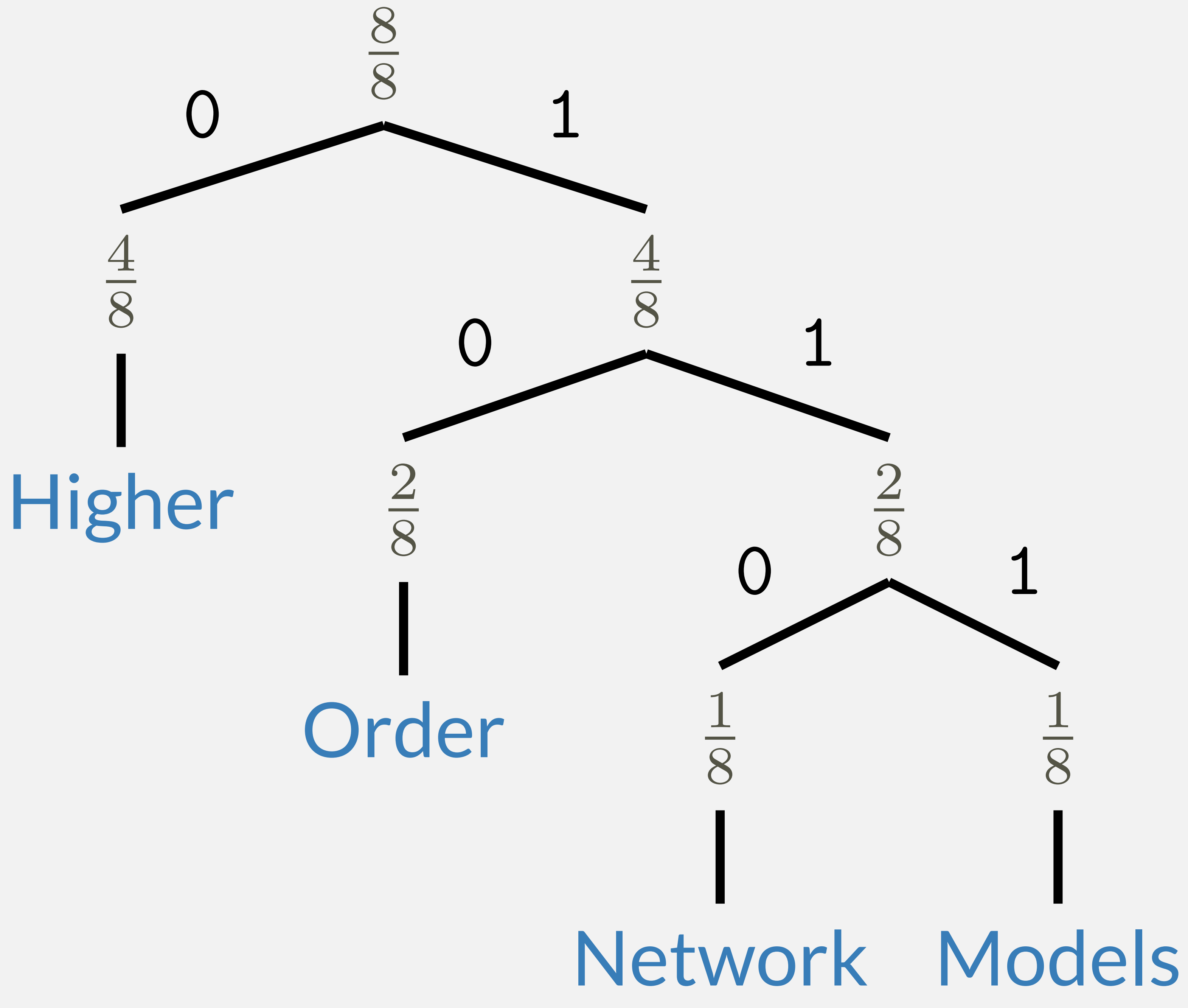
00 10 10 00 00 00 01 10 00 00 10 00 10 00 00 01 00 11 11 00 00 01 11 00
00 00 00 00 00 00 00 00 00 00 00 00 10 01 01 00 11 10 00 00 01 01 11 00
10 01 00 01 11 10 01 01 00 00 10 00 00 01 01 01 00 00 01 10 01 00 00 00
00 01 11 01 00 00 01 11 10 01 10 00 01 11 01 00 00 01 11 00 10 10 11 01
10 01 01



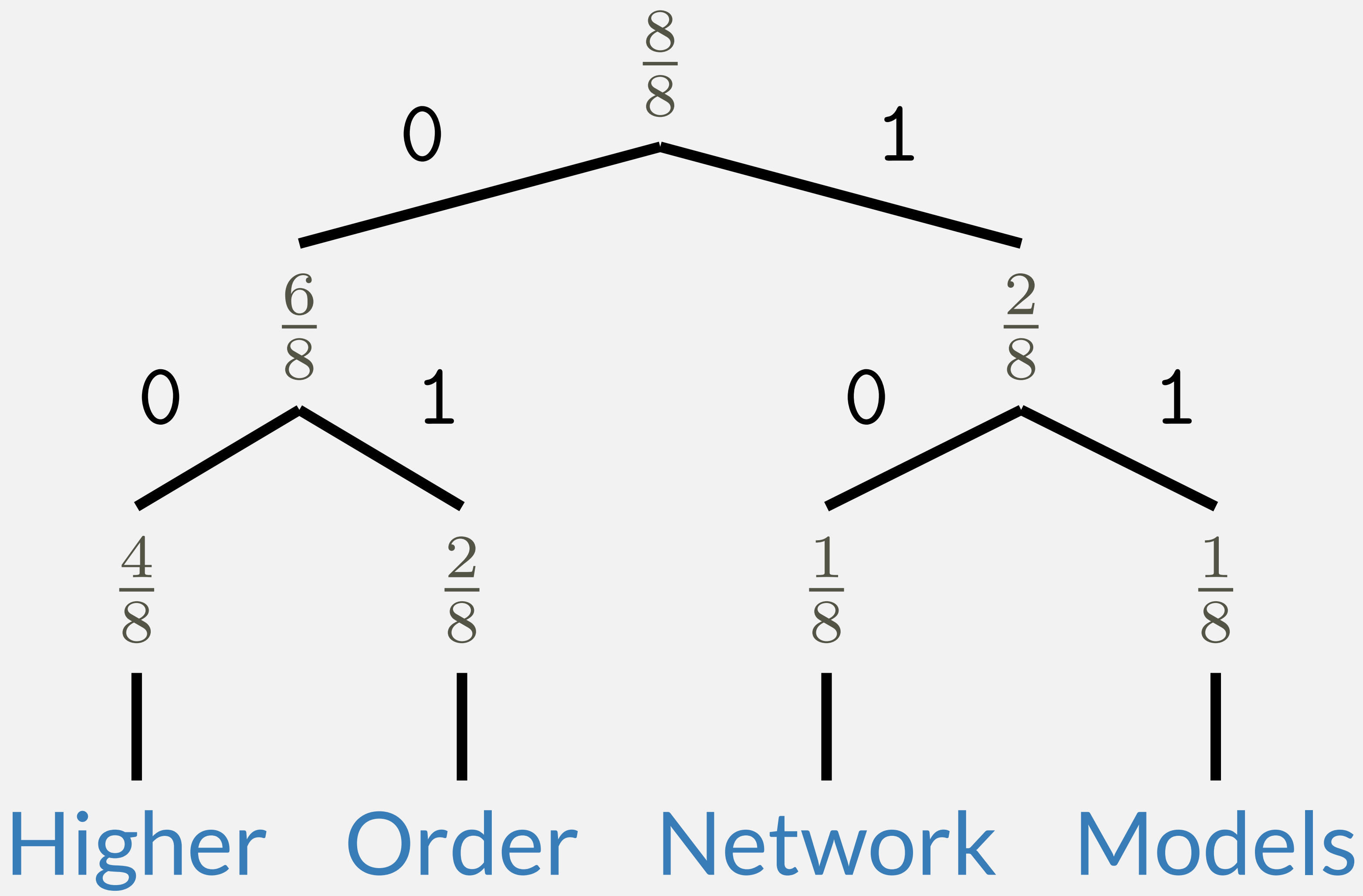
00101000000001100000100010000001001-
11100000011100000000000000000000000-
00100101001110000001011100100100011-
11001010000100000001010100001100100-
00000001110100000011110011000111010-
00001110010101101100101



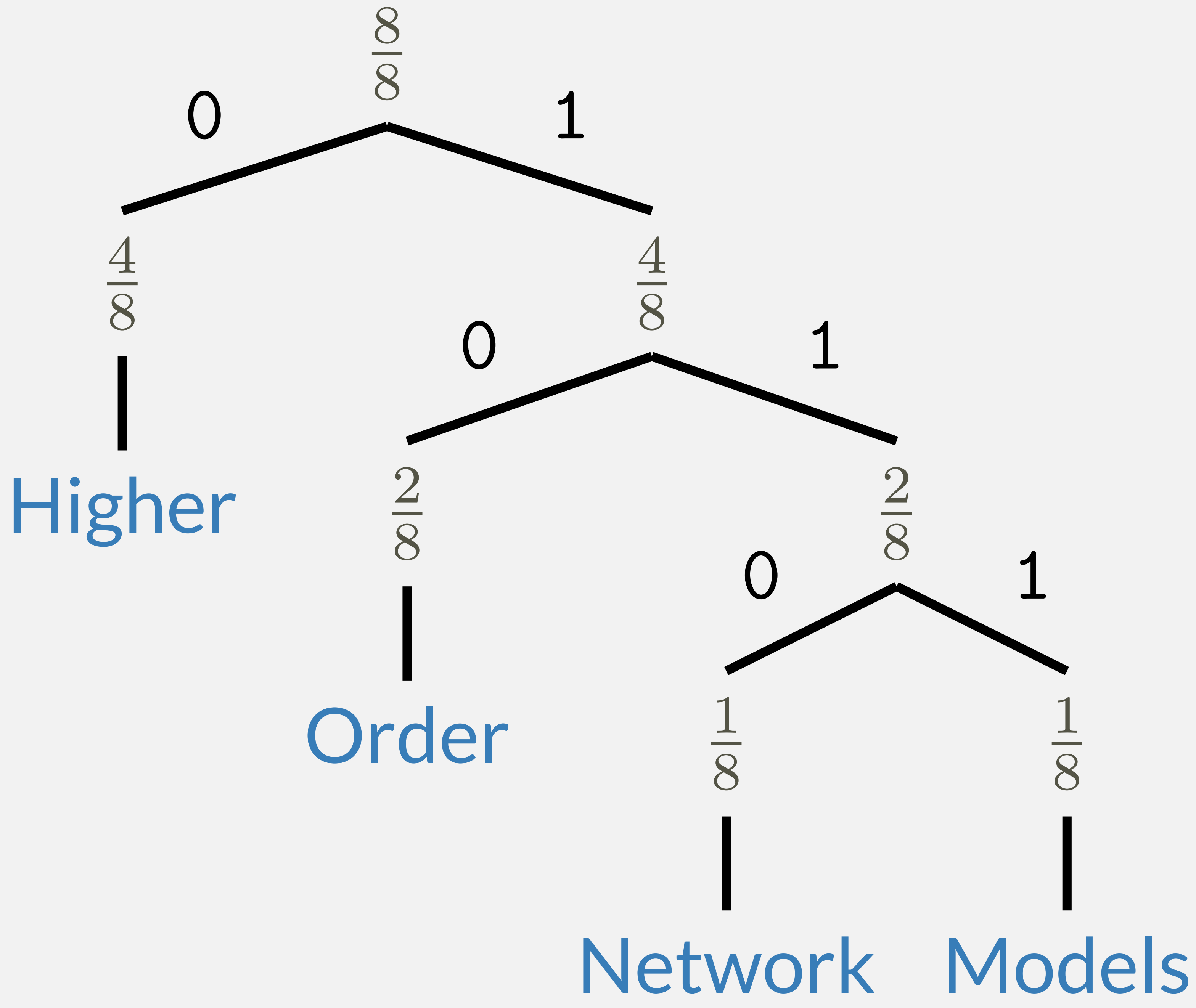
0010100000000110000010001000000 1001-
1110000011100000000000000000000000-
001001010011100000010111001001000 11-
1100101000010000000101010000001100100-
0000000111010000011110011000011 1010-
00001110010101101100101



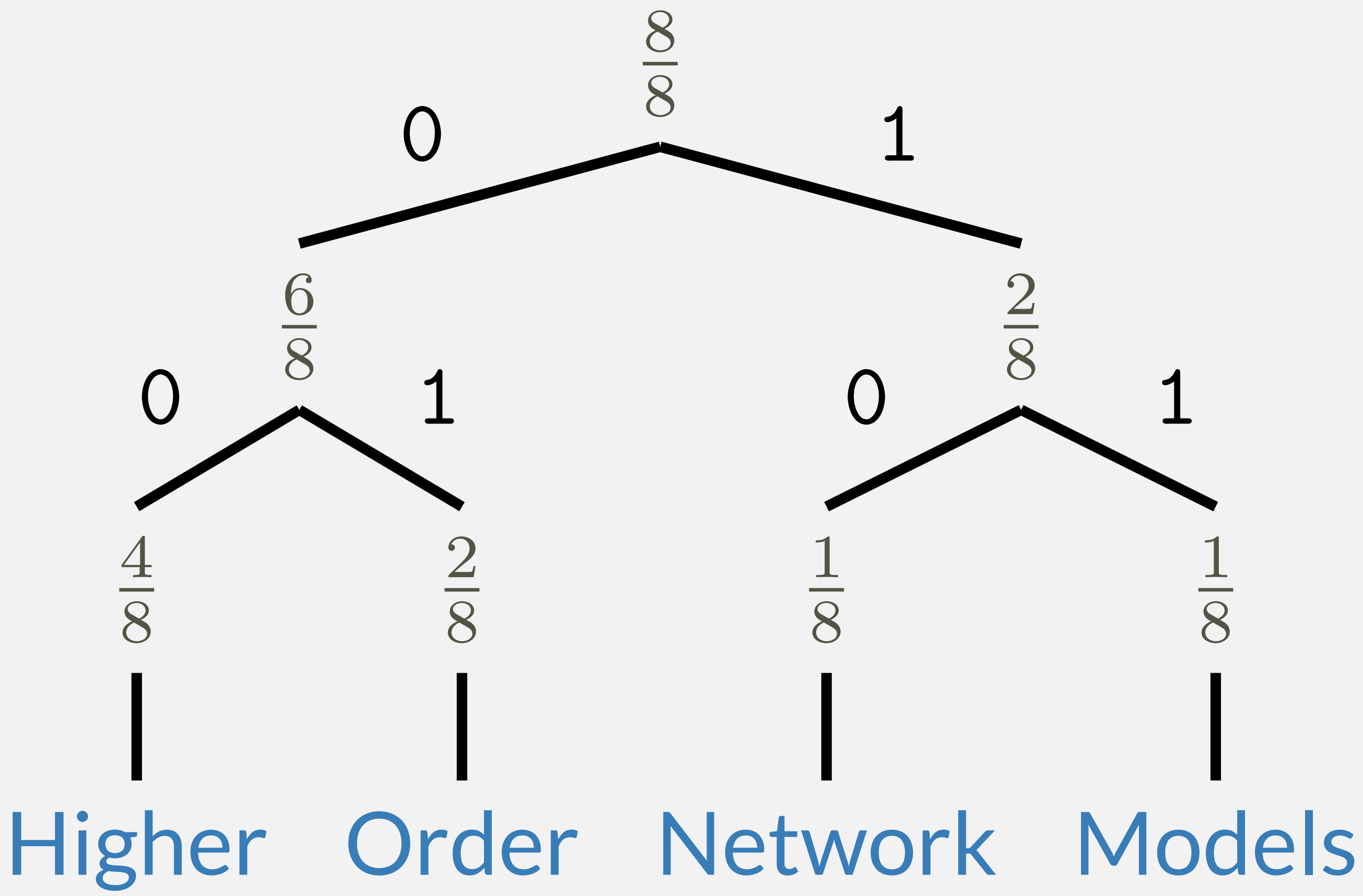
01101100001011000110011000100111111-
00101110000000000000110101001111100-
010101110110100101111101010000110001-
0101000101101000000101111000101111110-
10110010111100010111011011011110110-
1010



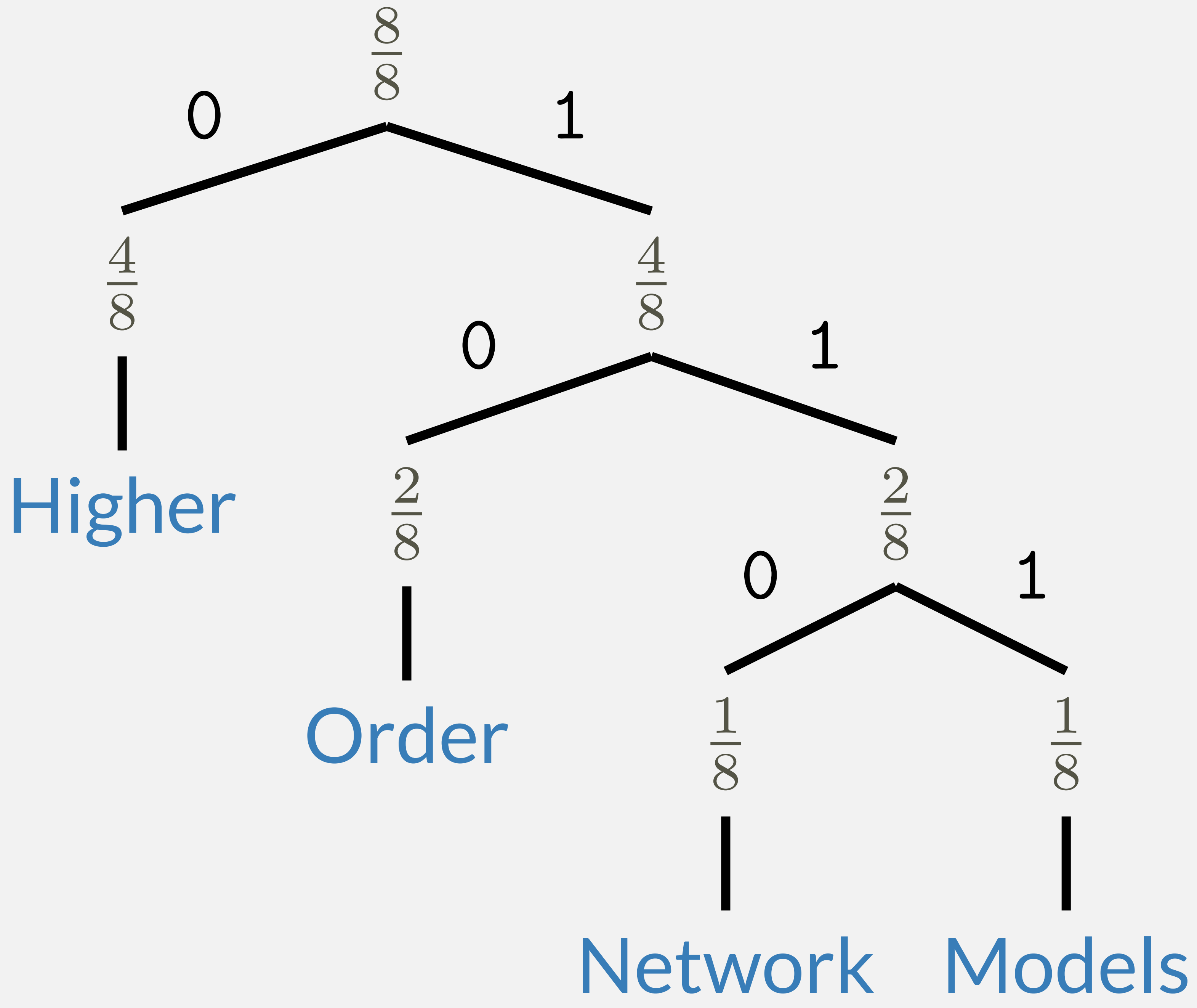
$$L = \sum_i p_i l_i = \frac{4}{8}2 + \frac{2}{8}2 + \frac{1}{8}2 + \frac{1}{8}2 = \mathbf{2}$$



$$L = \sum_i p_i l_i = \frac{4}{8}1 + \frac{2}{8}2 + \frac{1}{8}3 + \frac{1}{8}3 = \mathbf{\frac{7}{4}}$$

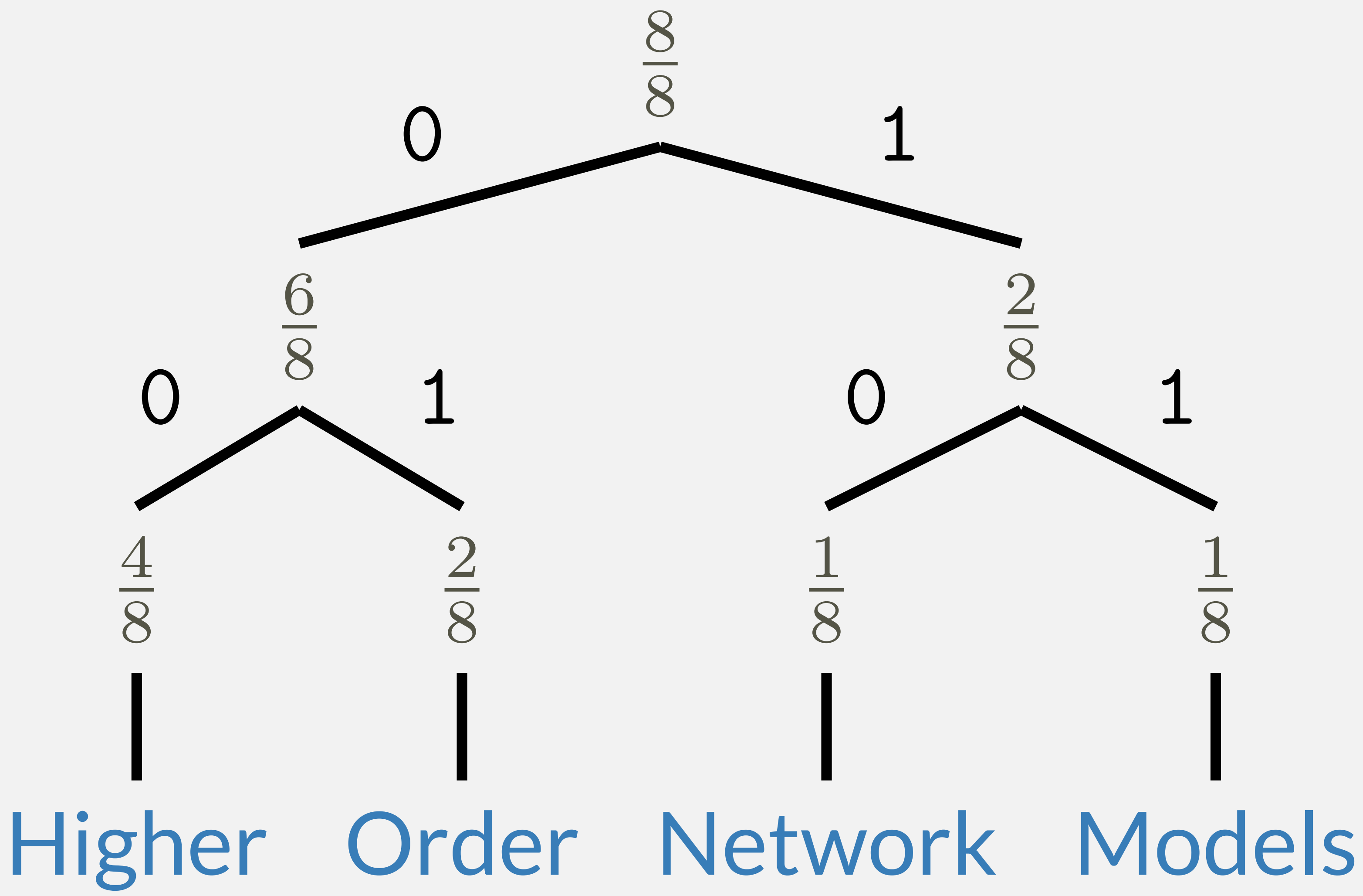


$$L = \sum_i p_i l_i = \frac{4}{8}2 + \frac{2}{8}2 + \frac{1}{8}2 + \frac{1}{8}2 = \mathbf{2}$$

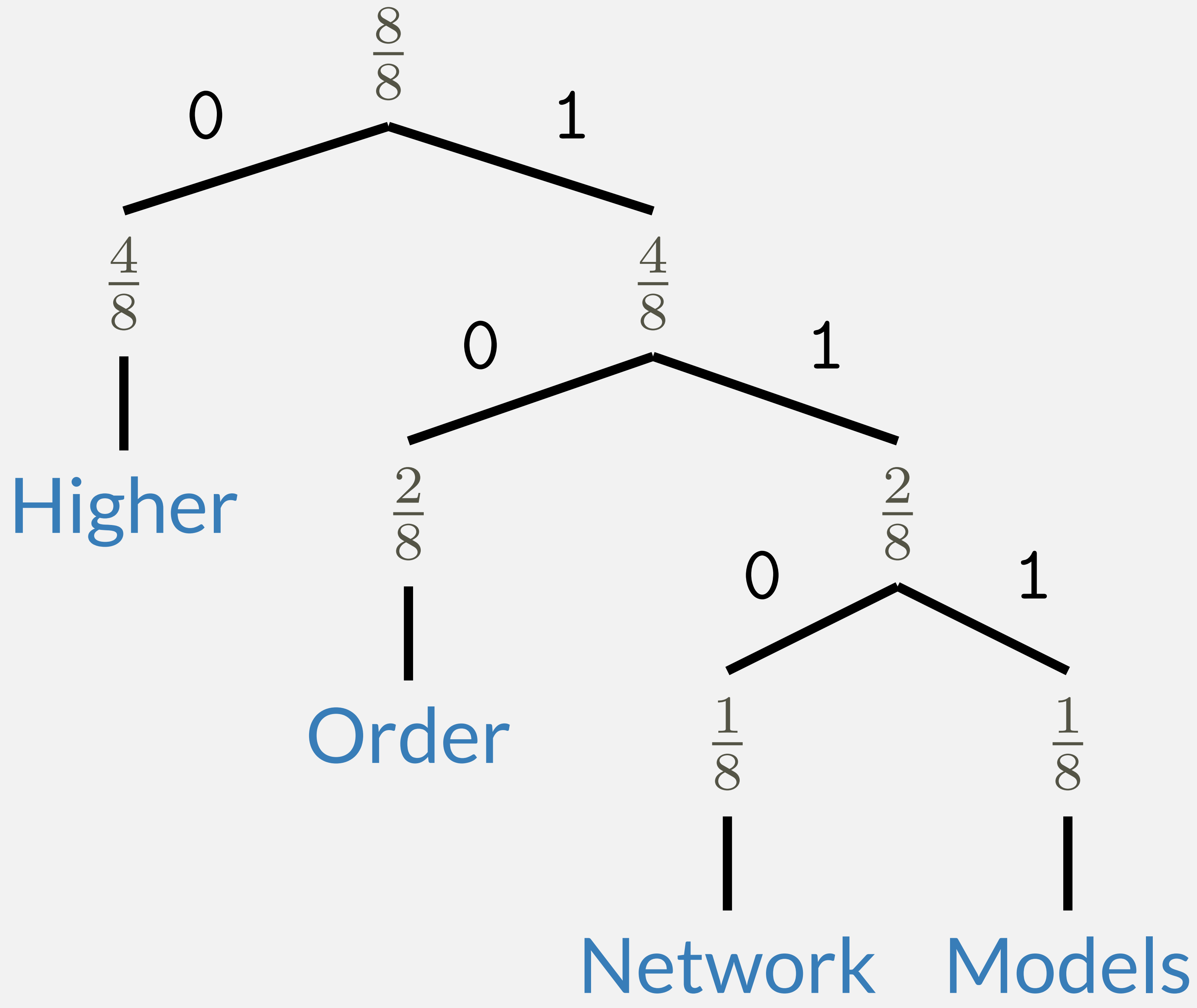


$$L = \sum_i p_i l_i = \frac{4}{8}1 + \frac{2}{8}2 + \frac{1}{8}3 + \frac{1}{8}3 = \mathbf{\frac{7}{4}}$$

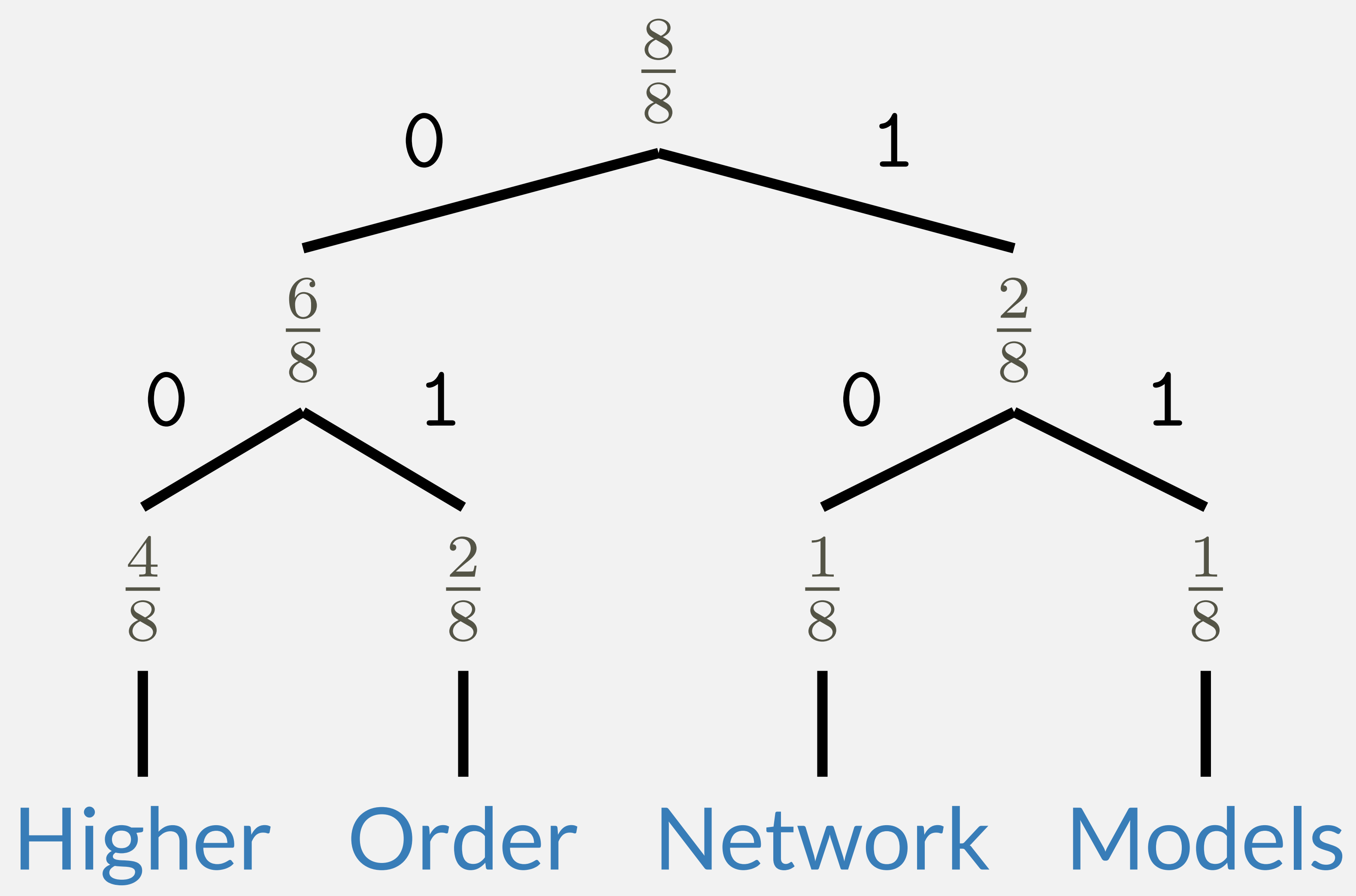
$$= \frac{4}{8}\log_2\frac{8}{4} + \frac{2}{8}\log_2\frac{8}{2} + \frac{1}{8}\log_2\frac{8}{1} + \frac{1}{8}\log_2\frac{8}{1}$$



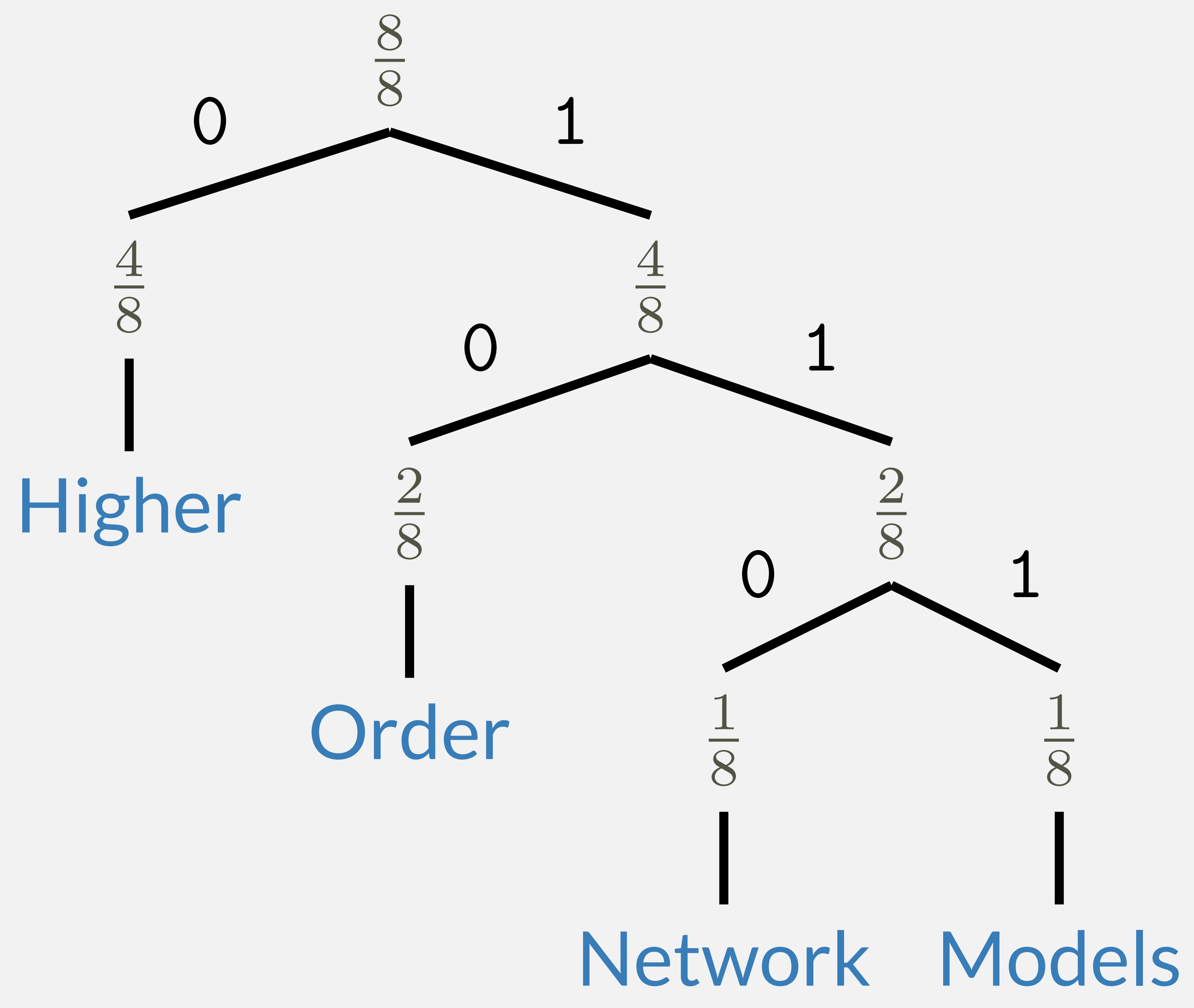
$$L = \sum_i p_i l_i = \frac{4}{8}2 + \frac{2}{8}2 + \frac{1}{8}2 + \frac{1}{8}2 = \mathbf{2}$$



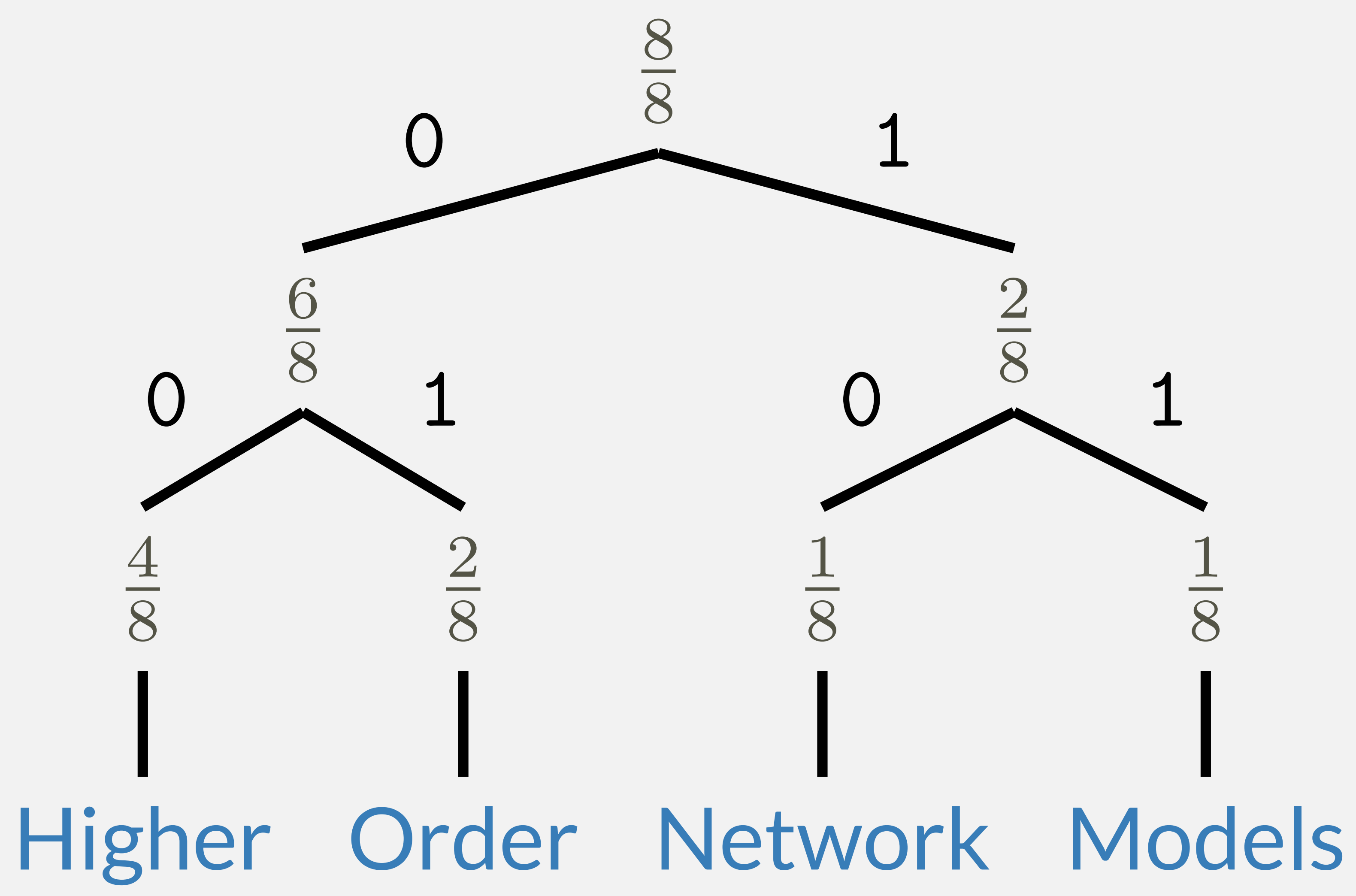
$$L = \sum_i p_i l_i = \frac{4}{8}1 + \frac{2}{8}2 + \frac{1}{8}3 + \frac{1}{8}3 = \mathbf{\frac{7}{4}}$$
$$= \frac{4}{8}\log_2\frac{8}{4} + \frac{2}{8}\log_2\frac{8}{2} + \frac{1}{8}\log_2\frac{8}{1} + \frac{1}{8}\log_2\frac{8}{1}$$
$$= -\frac{4}{8}\log_2\frac{4}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8}$$



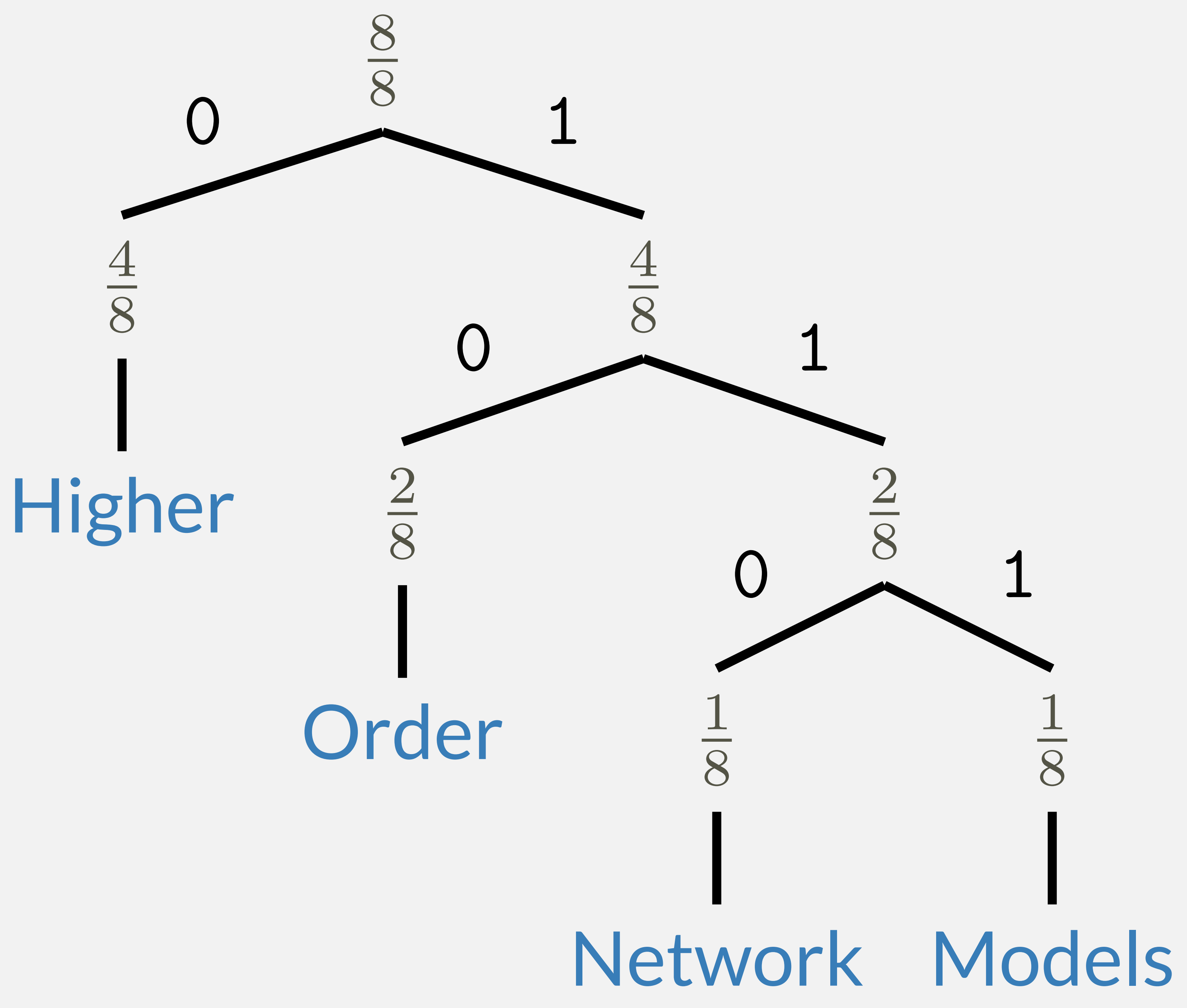
$$L = \sum_i p_i l_i = \frac{4}{8}2 + \frac{2}{8}2 + \frac{1}{8}2 + \frac{1}{8}2 = \mathbf{2}$$



$$\begin{aligned} L &= \sum_i p_i l_i = \frac{4}{8}1 + \frac{2}{8}2 + \frac{1}{8}3 + \frac{1}{8}3 = \mathbf{\frac{7}{4}} \\ &= \frac{4}{8}\log_2\frac{8}{4} + \frac{2}{8}\log_2\frac{8}{2} + \frac{1}{8}\log_2\frac{8}{1} + \frac{1}{8}\log_2\frac{8}{1} \\ &= -\frac{4}{8}\log_2\frac{4}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} \\ &= -\sum_i p_i \log_2 p_i \end{aligned}$$



$$L = \sum_i p_i l_i = \frac{4}{8}2 + \frac{2}{8}2 + \frac{1}{8}2 + \frac{1}{8}2 = \mathbf{2}$$



$$\begin{aligned} L &= \sum_i p_i l_i = \frac{4}{8}1 + \frac{2}{8}2 + \frac{1}{8}3 + \frac{1}{8}3 = \mathbf{\frac{7}{4}} \\ &= \frac{4}{8}\log_2\frac{8}{4} + \frac{2}{8}\log_2\frac{8}{2} + \frac{1}{8}\log_2\frac{8}{1} + \frac{1}{8}\log_2\frac{8}{1} \\ &= -\frac{4}{8}\log_2\frac{4}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} \\ &= -\sum_i p_i \log_2 p_i \\ &\equiv H(\mathcal{P}) \end{aligned}$$

1. **Coding theory**: The minimum description length principle

Regularities in data can be used to compress the data. The best compression captures most regularities

2. Mapping network flows:
The map equation

NETWORKS describe where flows move to depending on where they are

MAPS depict regularities using less information

If we can find a good code
for describing flows on a network,
we will have solved the dual problem
of finding the important structures
with respect to that flow

We use a modular code structure
that can exploit regions in the network
in which units of flow tend to stay
for a relatively long time

Two-level partitions

How many modules are present? And which nodes are members of which modules?

Maximal compression of flow with constraints:

1. Modular code structure
2. No more than two levels
3. Each node can only belong to one module

Two-level partitions with the map equation

How many modules are present? And which nodes are members of which modules?

Maximal compression of flow with constraints:

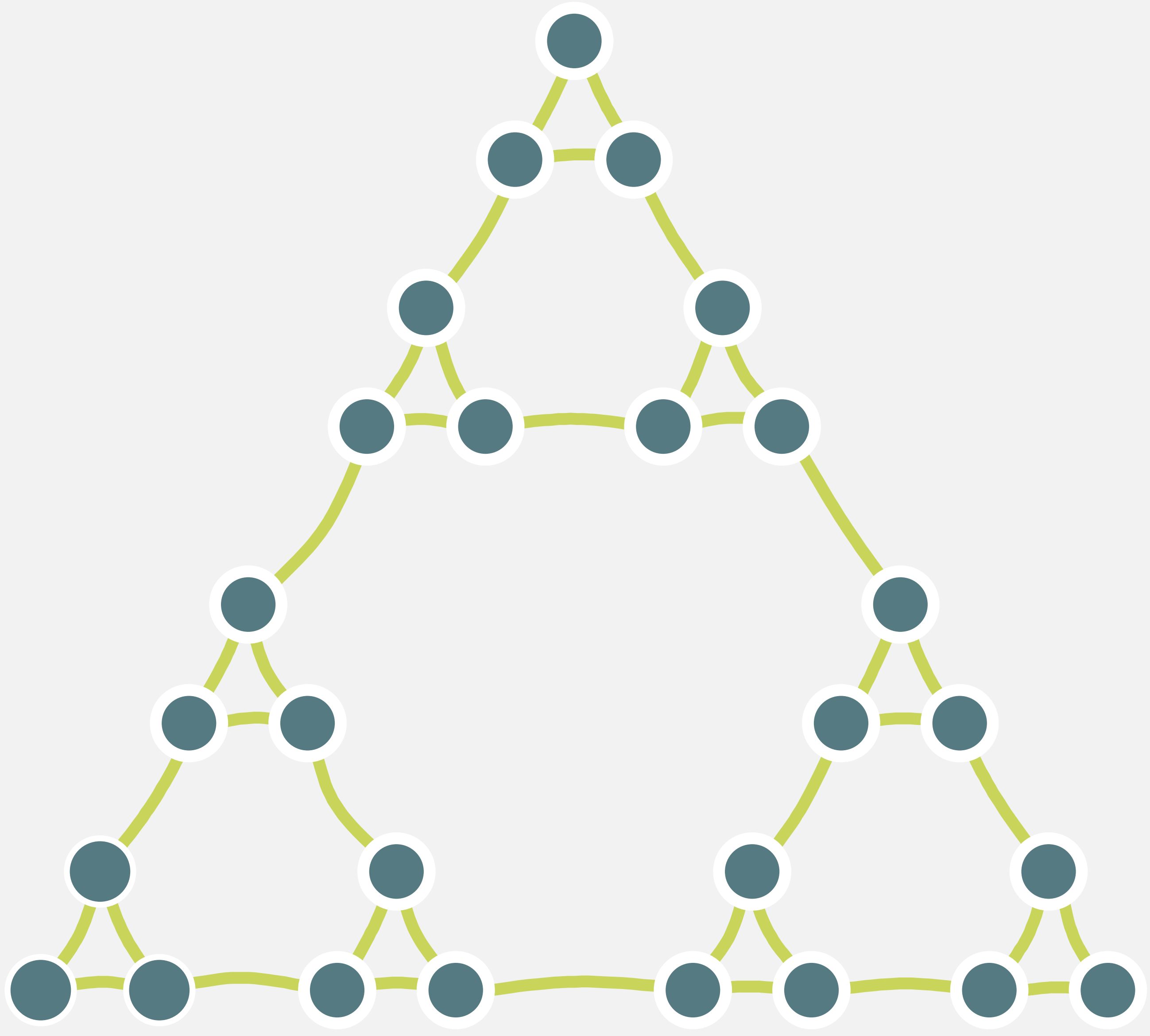
1. Modular code structure
2. No more than two levels
3. Each node can only belong to one module

Two-level partitions with the map equation

$$L(\mathsf{M}) = q_{\curvearrowleft} H(\mathcal{Q}) + \sum_{i=1}^{m} p_{\circlearrowright}^{i} H(\mathcal{P}^{i})$$

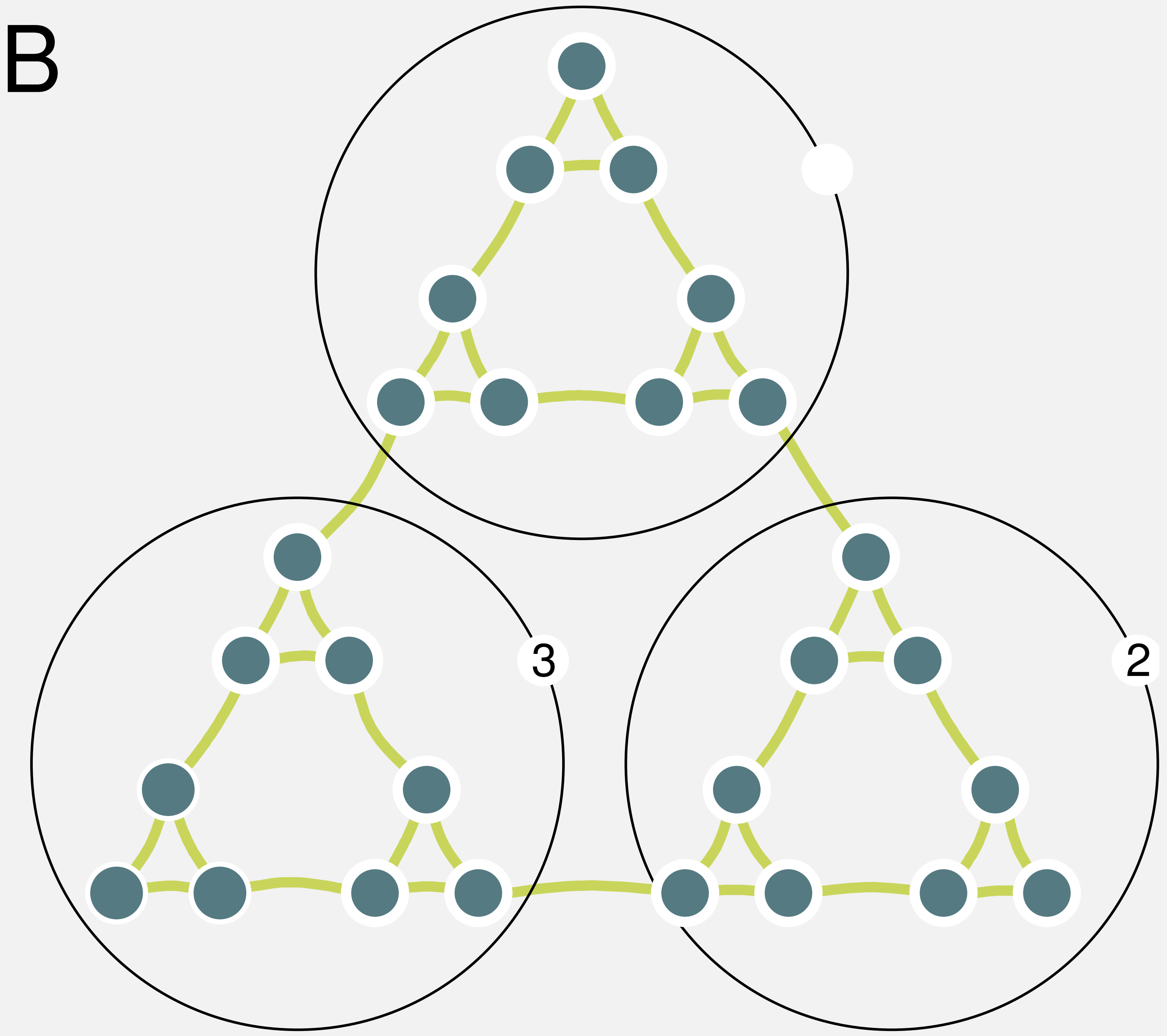
Two-level partitions with the map equation

A



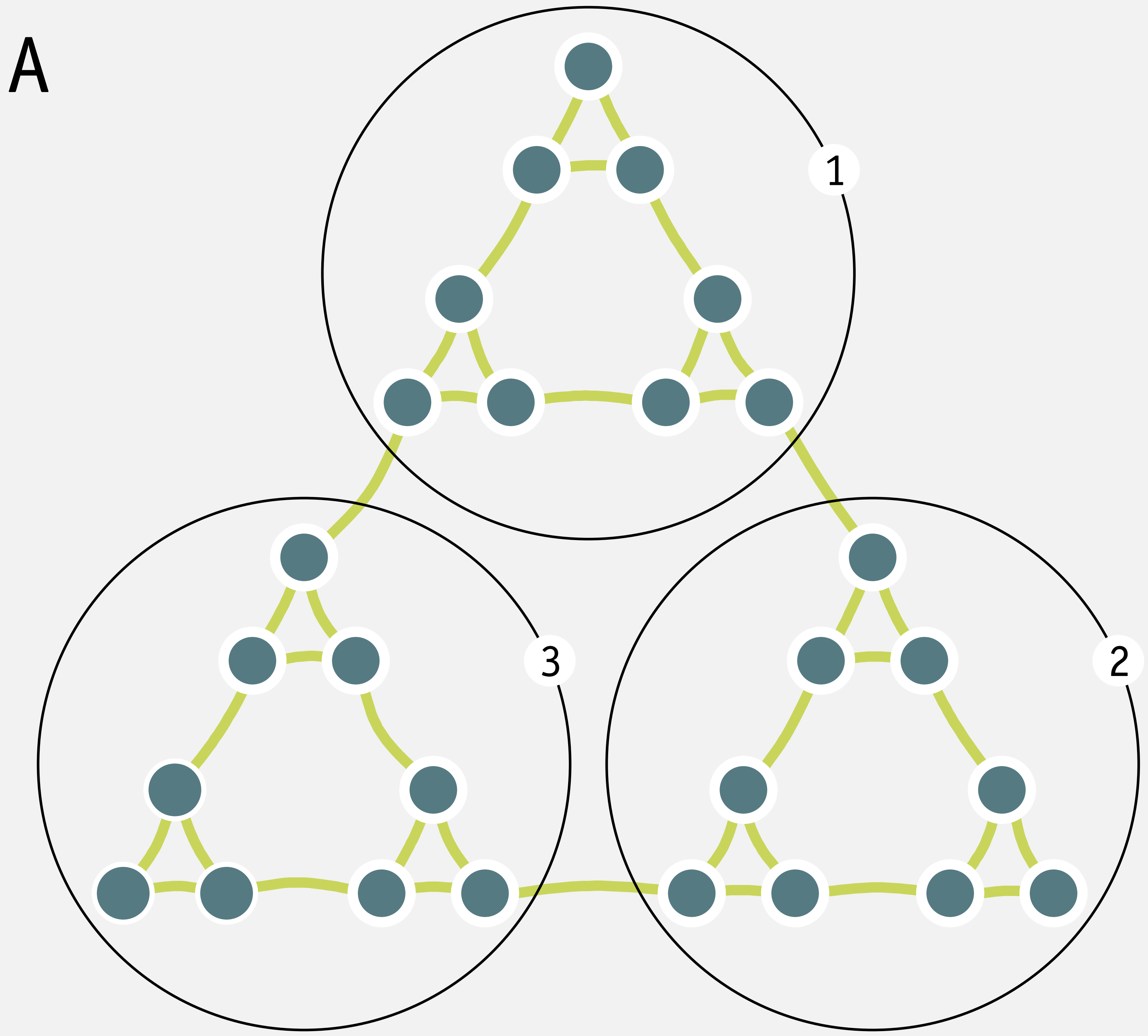
$$L(\mathsf{M}) = H(\mathcal{P}) = 4.75 \text{ bits.}$$

B

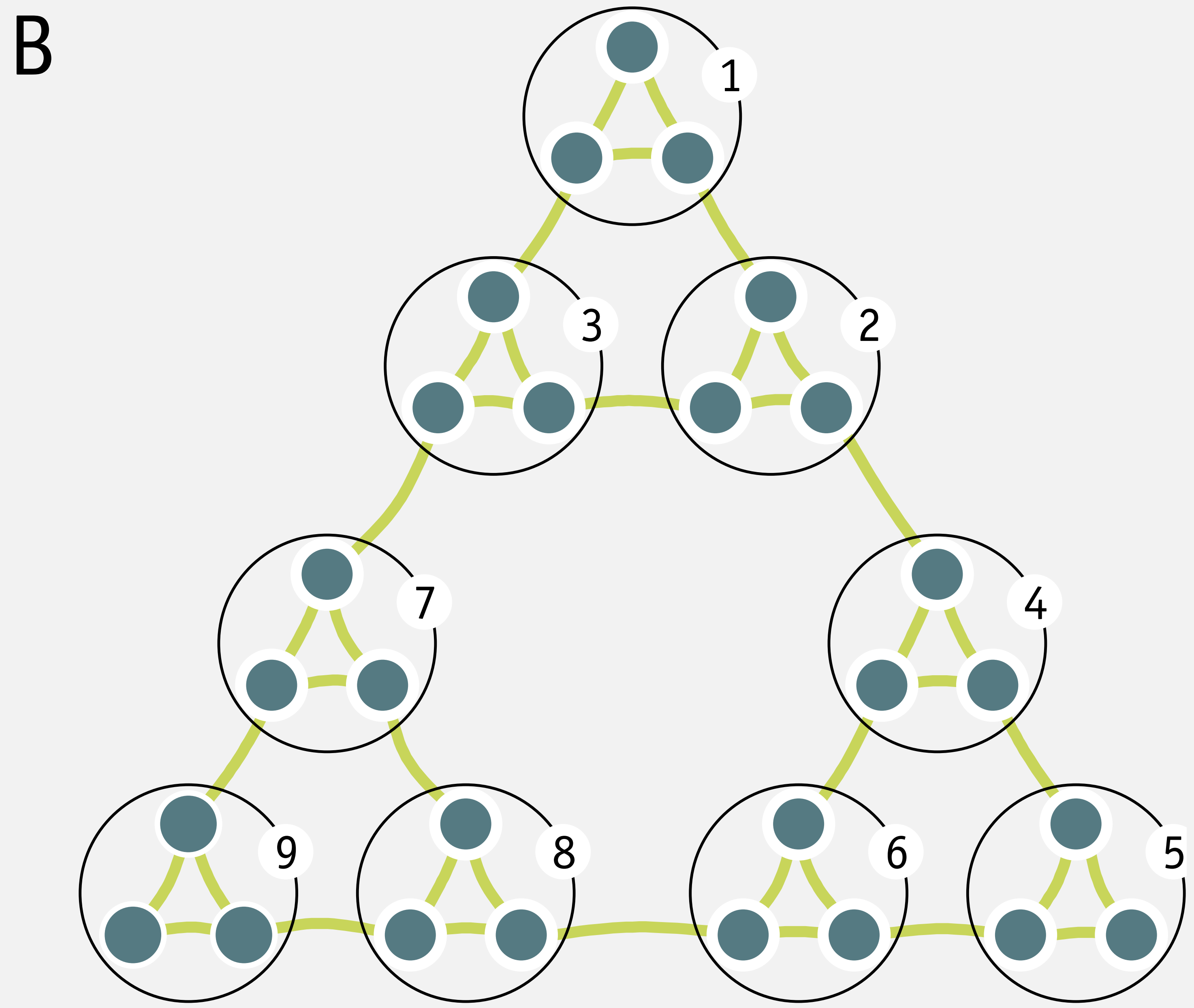


$$L(\mathsf{M}) = \underbrace{q_{\curvearrowleft} H(\mathcal{Q})}_{0.12 \text{ bits}} + \underbrace{\begin{cases} p^1_{\circlearrowright} H(\mathcal{P}^1) \\ p^2_{\circlearrowright} H(\mathcal{P}^2) \\ p^3_{\circlearrowright} H(\mathcal{P}^3) \end{cases}}_{3.56 \text{ bits}} = 3.68 \text{ bits.}$$

Two-level partitions with the map equation



$$L(\mathsf{M}) = \underbrace{q_{\curvearrowright} H(\mathcal{Q})}_{0.12\ \text{bits}} + \underbrace{\begin{cases} p_{\circlearrowright}^{1} H(\mathcal{P}^{1}) \\ p_{\circlearrowright}^{2} H(\mathcal{P}^{2}) \\ p_{\circlearrowright}^{3} H(\mathcal{P}^{3}) \end{cases}}_{3.56\ \text{bits}} = 3.68\ \text{bits}.$$

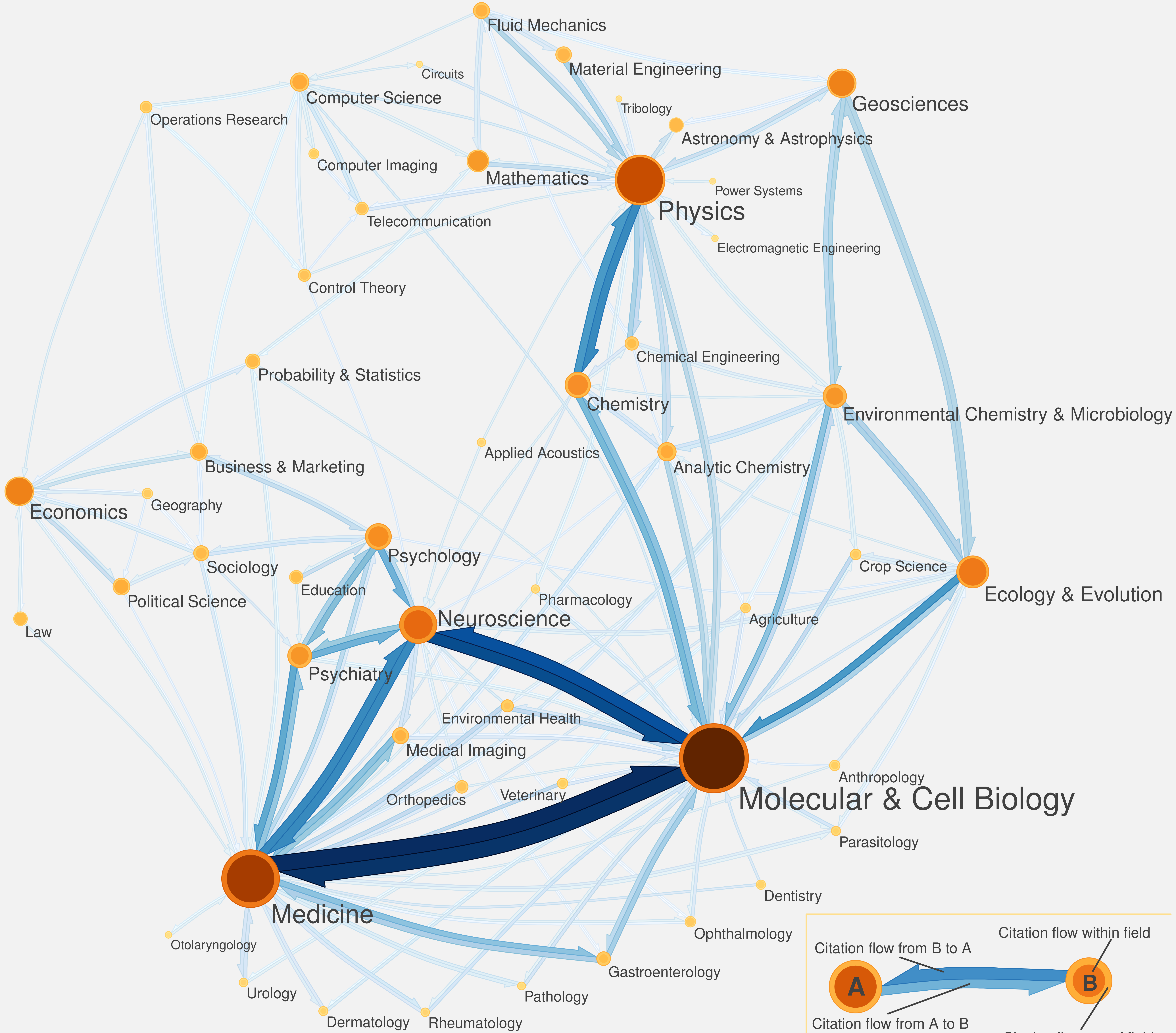


$$L(\mathsf{M}) = \underbrace{q_{\curvearrowright} H(\mathcal{Q})}_{0.97\ \text{bits}} + \underbrace{\begin{cases} p_{\circlearrowright}^{1} H(\mathcal{P}^{1}) \\ p_{\circlearrowright}^{2} H(\mathcal{P}^{2}) \\ p_{\circlearrowright}^{3} H(\mathcal{P}^{3}) \\ p_{\circlearrowright}^{4} H(\mathcal{P}^{4}) \\ p_{\circlearrowright}^{5} H(\mathcal{P}^{5}) \\ p_{\circlearrowright}^{6} H(\mathcal{P}^{6}) \\ p_{\circlearrowright}^{7} H(\mathcal{P}^{7}) \\ p_{\circlearrowright}^{8} H(\mathcal{P}^{8}) \\ p_{\circlearrowright}^{9} H(\mathcal{P}^{9}) \end{cases}}_{2.60\ \text{bits}} = 3.57\ \text{bits}.$$

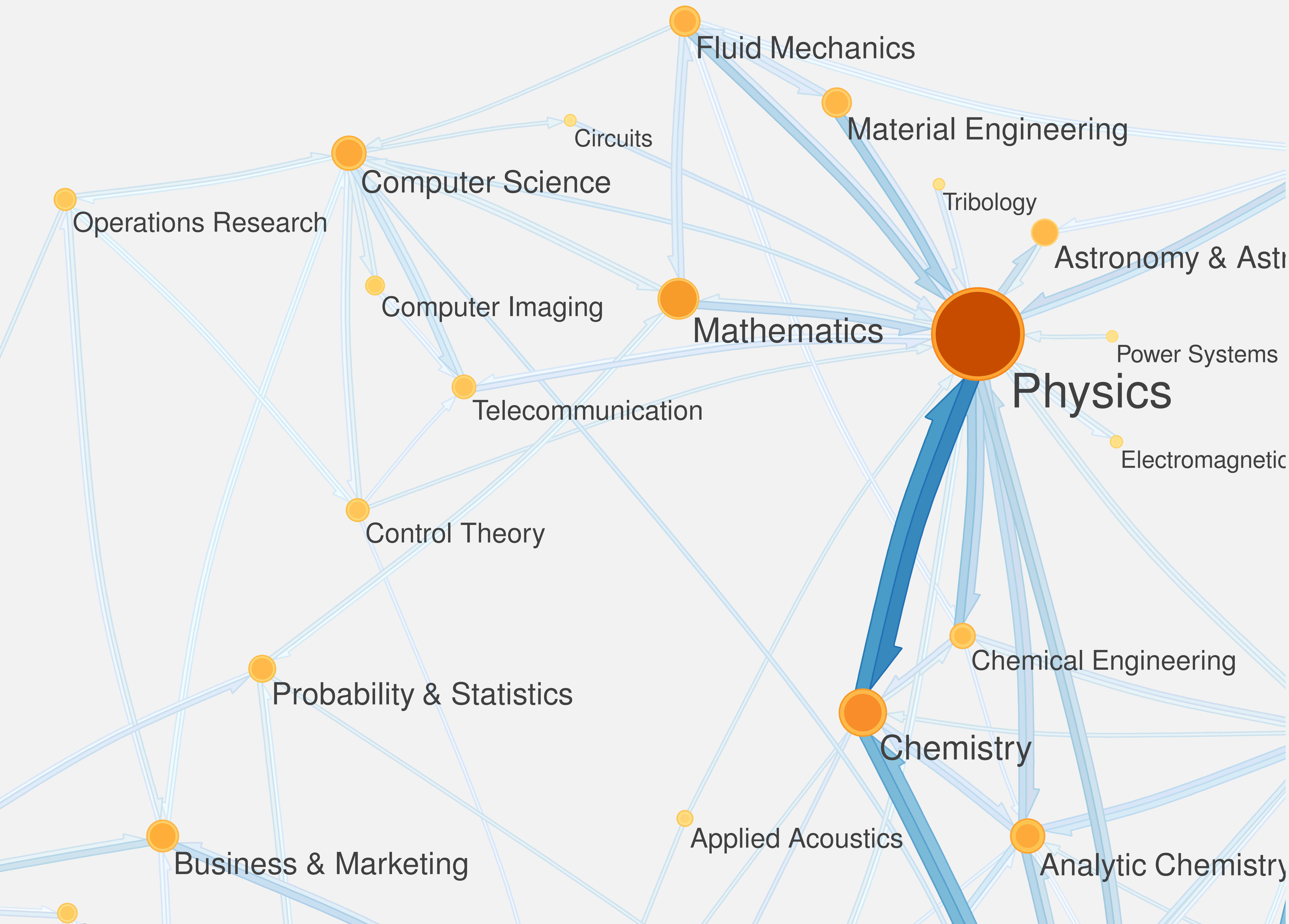
Science 2010

10,000 journals
1,000,000 articles
10,000,000 citations

Thomson Scientific Journal Citation Reports
2010



Fluid Mechanics
Material Engineering
Circuits
Computer Science
Operations Research
Tribology
Geosciences
Astronomy & Astrophysics
Computer Imaging
Mathematics
Power Systems
Physics
Telecommunication
Electromagnetic Engineering
Control Theory
Probability & Statistics
Chemical Engineering
Chemistry
Environmental Chemistry & Microbiology
Applied Acoustics
Analytic Chemistry
Business & Marketing
Economics
Geography
Psychology
Sociology
Crop Science
Ecology & Evolution
Education
Political Science
Pharmacology
Law
Neuroscience
Agriculture
Psychiatry
Environmental Health
Medical Imaging
Anthropology
Orthopedics
Veterinary
Molecular & Cell Biology
Parasitology
Medicine
Dentistry
Otolaryngology
Ophthalmology
Gastroenterology
Urology
Pathology
Dermatology
Rheumatology
Citation flow within field
Citation flow from B to A
A
B
Citation flow from A to B



Fluid Mechanics
Material Engineering
Circuits
Computer Science
Operations Research
Tribology
Astronomy & Astr
Computer Imaging
Mathematics
Power Systems
Physics
Telecommunication
Electromagnetic
Control Theory
Chemical Engineering
Probability & Statistics
Chemistry
Applied Acoustics
Analytic Chemistry
Business & Marketing

Multilevel partitions

Into how many hierarchical levels is a given network organized? How many modules are present at each level? And which nodes are members of which modules?

Maximal compression of flow with constraints:

1. Modular code structure
2. ~~No more than two levels~~
3. Each node can only belong to one module

Multilevel partitions with the map equation

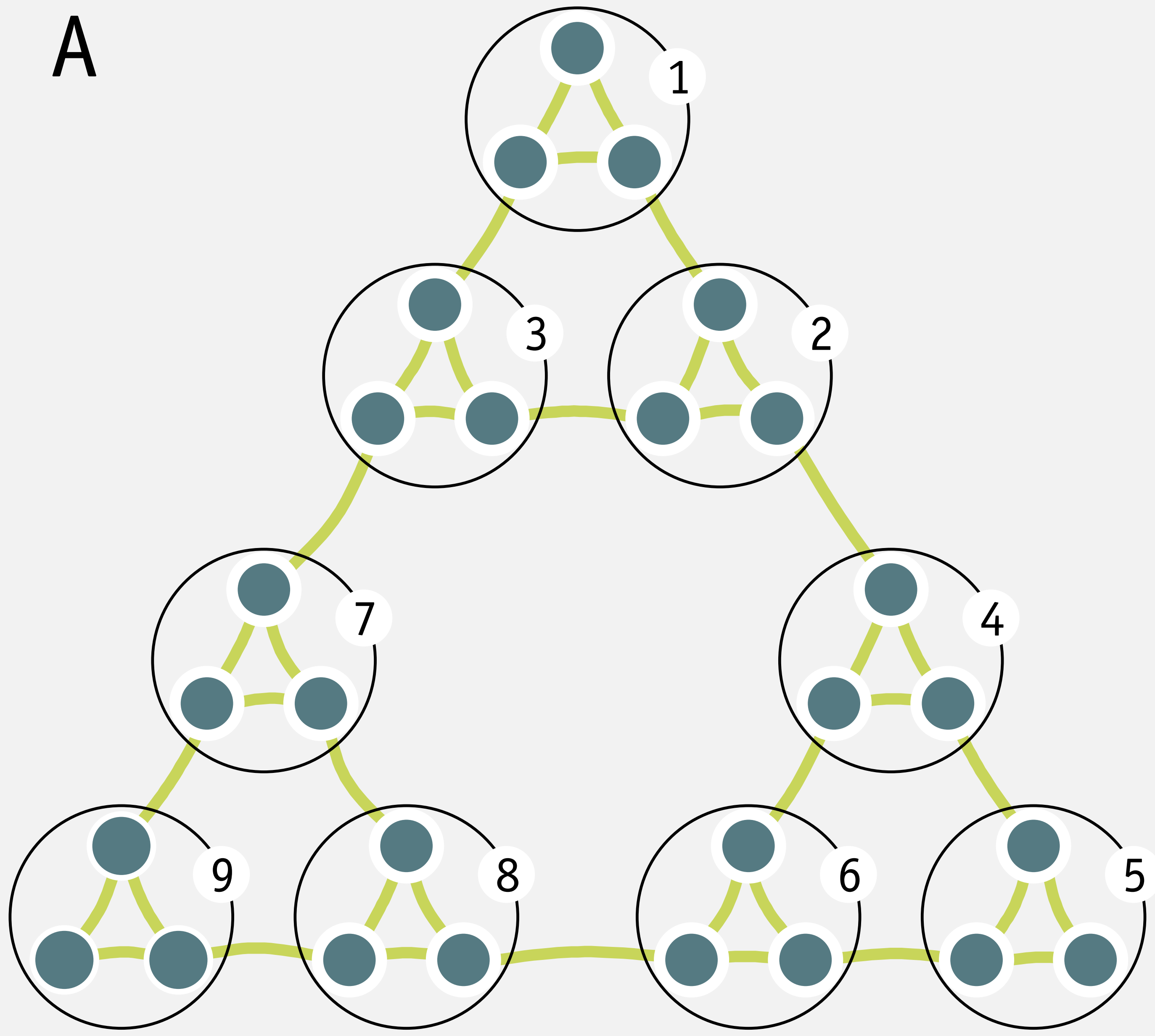
Into how many hierarchical levels is a given network organized? How many modules are present at each level? And which nodes are members of which modules?

Maximal compression of flow with constraints:

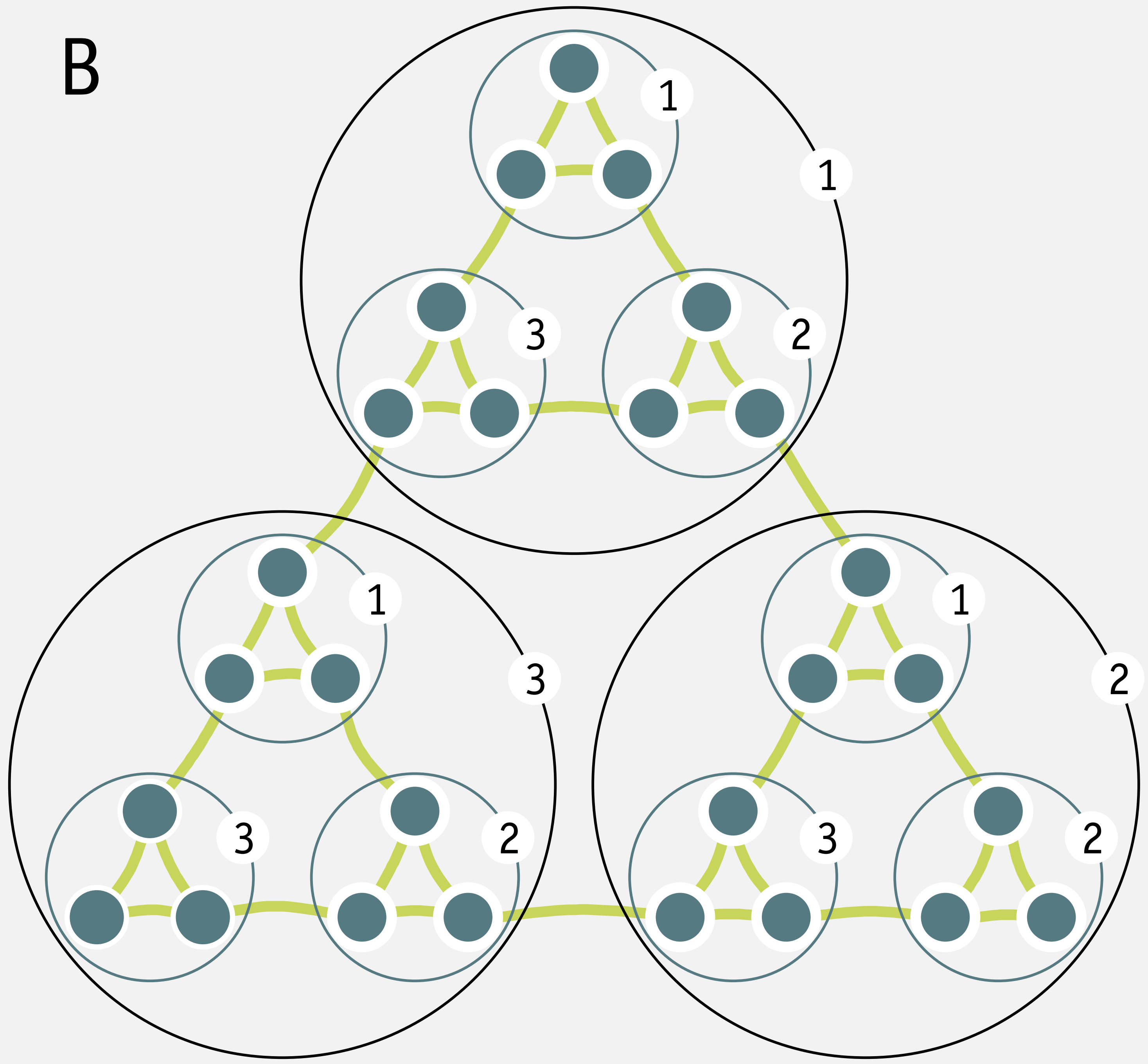
1. Modular code structure
2. ~~No more than two levels~~
3. Each node can only belong to one module

Multilevel partitions with the map equation

A



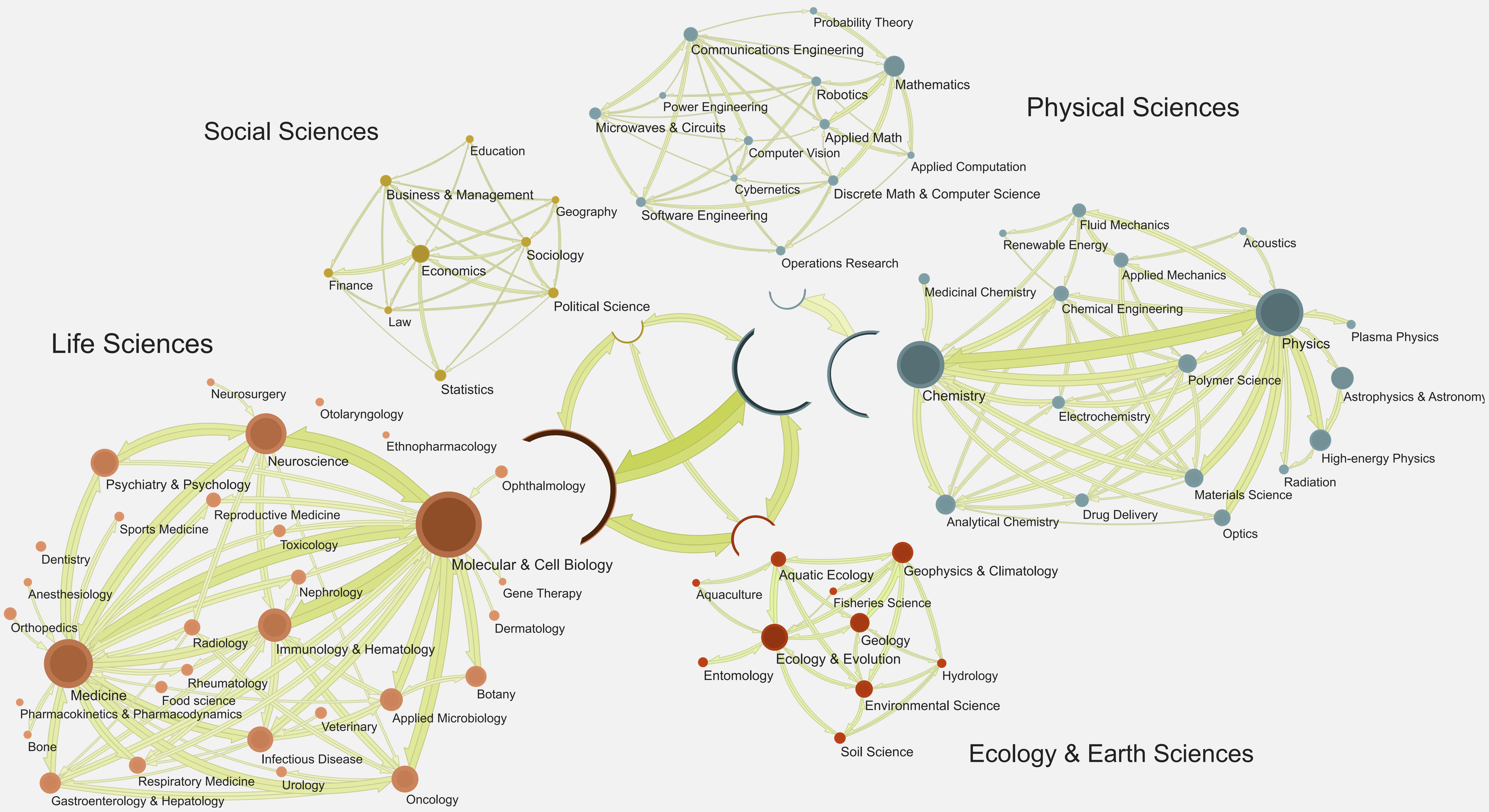
B



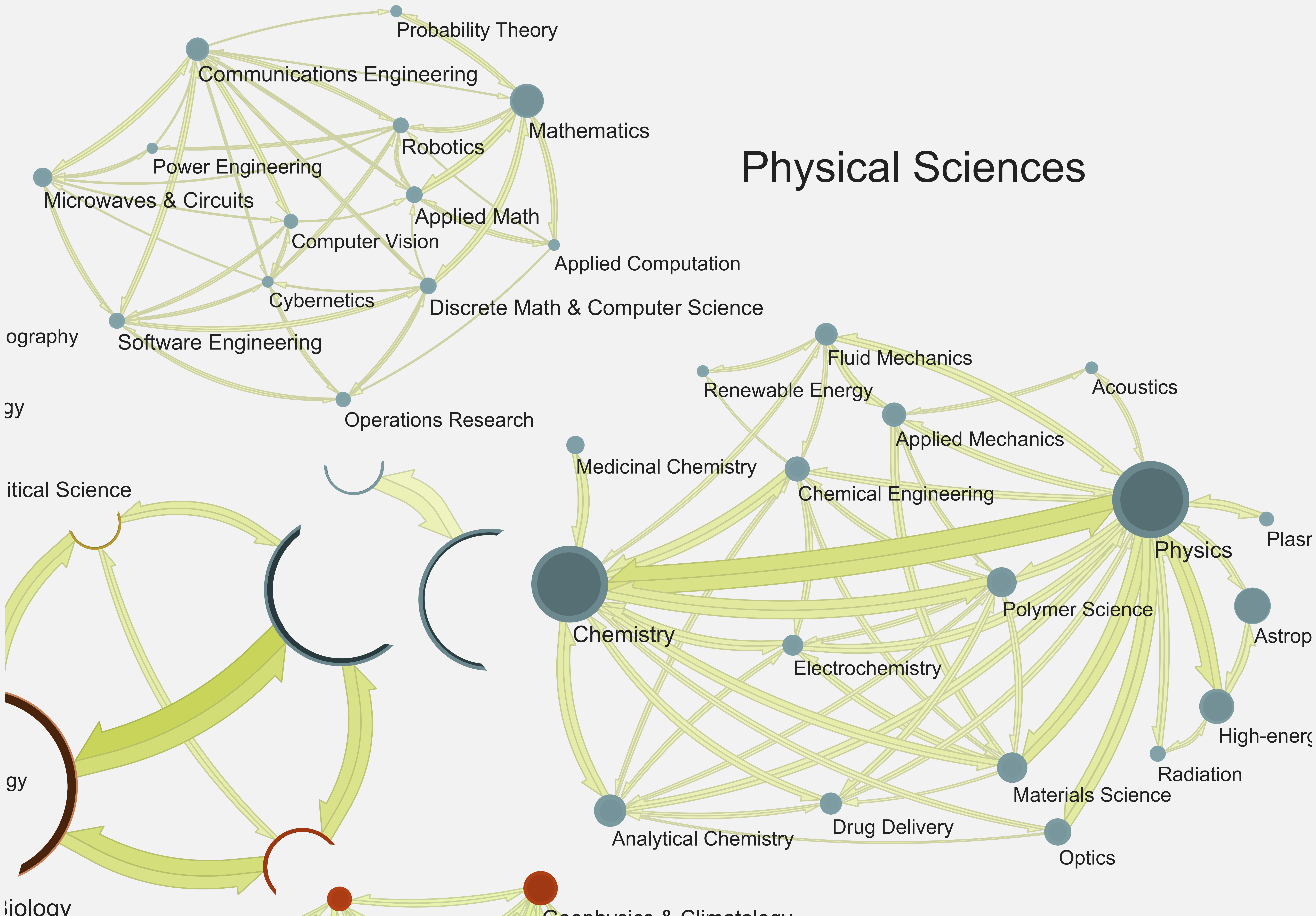
$$L(\mathsf{M}) = \underbrace{q_{\curvearrowleft} H(\mathcal{Q})}_{0.97\text{ bits}} + \underbrace{\begin{cases} p^{1}_{\circlearrowright} H(\mathcal{P}^{1}) \\ p^{2}_{\circlearrowright} H(\mathcal{P}^{2}) \\ p^{3}_{\circlearrowright} H(\mathcal{P}^{3}) \\ p^{4}_{\circlearrowright} H(\mathcal{P}^{4}) \\ p^{5}_{\circlearrowright} H(\mathcal{P}^{5}) \\ p^{6}_{\circlearrowright} H(\mathcal{P}^{6}) \\ p^{7}_{\circlearrowright} H(\mathcal{P}^{7}) \\ p^{8}_{\circlearrowright} H(\mathcal{P}^{8}) \\ p^{9}_{\circlearrowright} H(\mathcal{P}^{9}) \end{cases}}_{2.60\text{ bits}} = 3.57\text{ bits.}$$

$$L(\mathsf{M}) = q_{\curvearrowleft} H(\mathcal{Q}) + \begin{cases} q^{1}_{\circlearrowright} H(\mathcal{Q}^{1}) + \begin{cases} p^{11}_{\circlearrowright} H(\mathcal{P}^{11}) \\ p^{12}_{\circlearrowright} H(\mathcal{P}^{12}) \\ p^{13}_{\circlearrowright} H(\mathcal{P}^{13}) \end{cases} \\ q^{2}_{\circlearrowright} H(\mathcal{Q}^{2}) + \begin{cases} p^{21}_{\circlearrowright} H(\mathcal{P}^{21}) \\ p^{22}_{\circlearrowright} H(\mathcal{P}^{22}) \\ p^{23}_{\circlearrowright} H(\mathcal{P}^{23}) \end{cases} \\ q^{3}_{\circlearrowright} H(\mathcal{Q}^{3}) + \begin{cases} p^{31}_{\circlearrowright} H(\mathcal{P}^{31}) \\ p^{32}_{\circlearrowright} H(\mathcal{P}^{32}) \\ p^{33}_{\circlearrowright} H(\mathcal{P}^{33}) \end{cases} \end{cases} = 3.48\text{ bits.}$$

0.12 bits   0.76 bits   2.60 bits



Social Sciences
Education
Business & Management
Geography
Sociology
Economics
Finance
Political Science
Law
Statistics
Physical Sciences
Probability Theory
Communications Engineering
Mathematics
Robotics
Power Engineering
Microwaves & Circuits
Applied Math
Computer Vision
Applied Computation
Cybernetics
Discrete Math & Computer Science
Software Engineering
Operations Research
Fluid Mechanics
Renewable Energy
Acoustics
Applied Mechanics
Medicinal Chemistry
Chemical Engineering
Plasma Physics
Physics
Chemistry
Polymer Science
Astrophysics & Astronomy
Electrochemistry
High-energy Physics
Radiation
Materials Science
Drug Delivery
Analytical Chemistry
Optics
Life Sciences
Neurosurgery
Otolaryngology
Ethnopharmacology
Neuroscience
Psychiatry & Psychology
Ophthalmology
Reproductive Medicine
Sports Medicine
Toxicology
Dentistry
Molecular & Cell Biology
Anesthesiology
Nephrology
Gene Therapy
Orthopedics
Dermatology
Radiology
Immunology & Hematology
Rheumatology
Medicine
Food science
Botany
Pharmacokinetics & Pharmacodynamics
Applied Microbiology
Veterinary
Bone
Infectious Disease
Respiratory Medicine
Urology
Gastroenterology & Hepatology
Oncology
Ecology & Earth Sciences
Aquatic Ecology
Geophysics & Climatology
Aquaculture
Fisheries Science
Geology
Ecology & Evolution
Entomology
Hydrology
Environmental Science
Soil Science



Physical Sciences
Probability Theory
Communications Engineering
Mathematics
Robotics
Power Engineering
Microwaves & Circuits
Applied Math
Computer Vision
Applied Computation
Cybernetics
Discrete Math & Computer Science
Software Engineering
Operations Research
Fluid Mechanics
Renewable Energy
Acoustics
Applied Mechanics
Medicinal Chemistry
Chemical Engineering
Physics
Chemistry
Polymer Science
Electrochemistry
Radiation
Materials Science
Drug Delivery
Analytical Chemistry
Optics

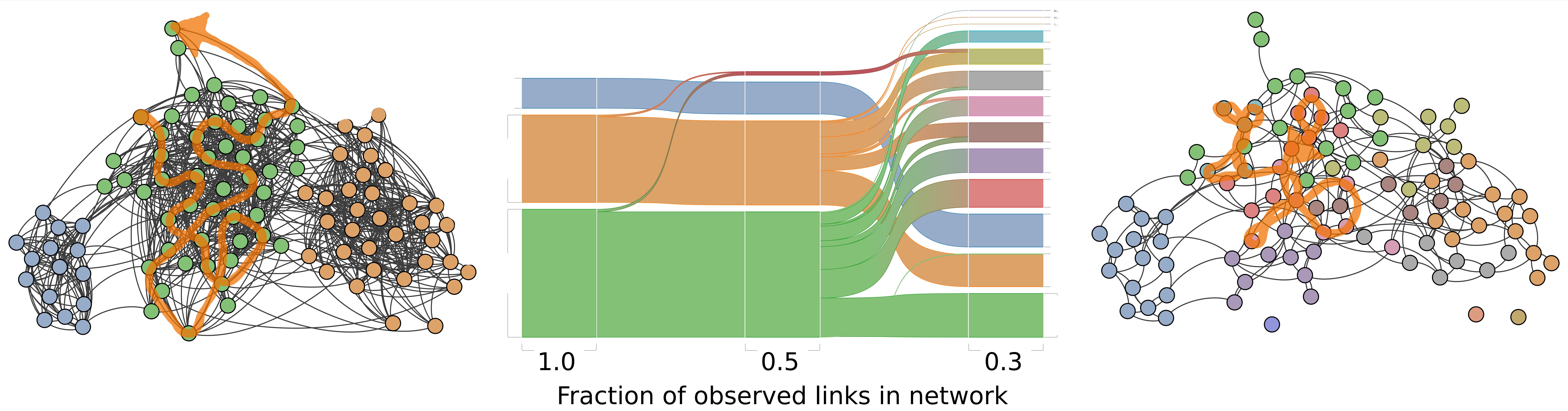
2. Mapping network flows:
The map equation

The map equation infers communities with long flow persistence using the minimum description length principle

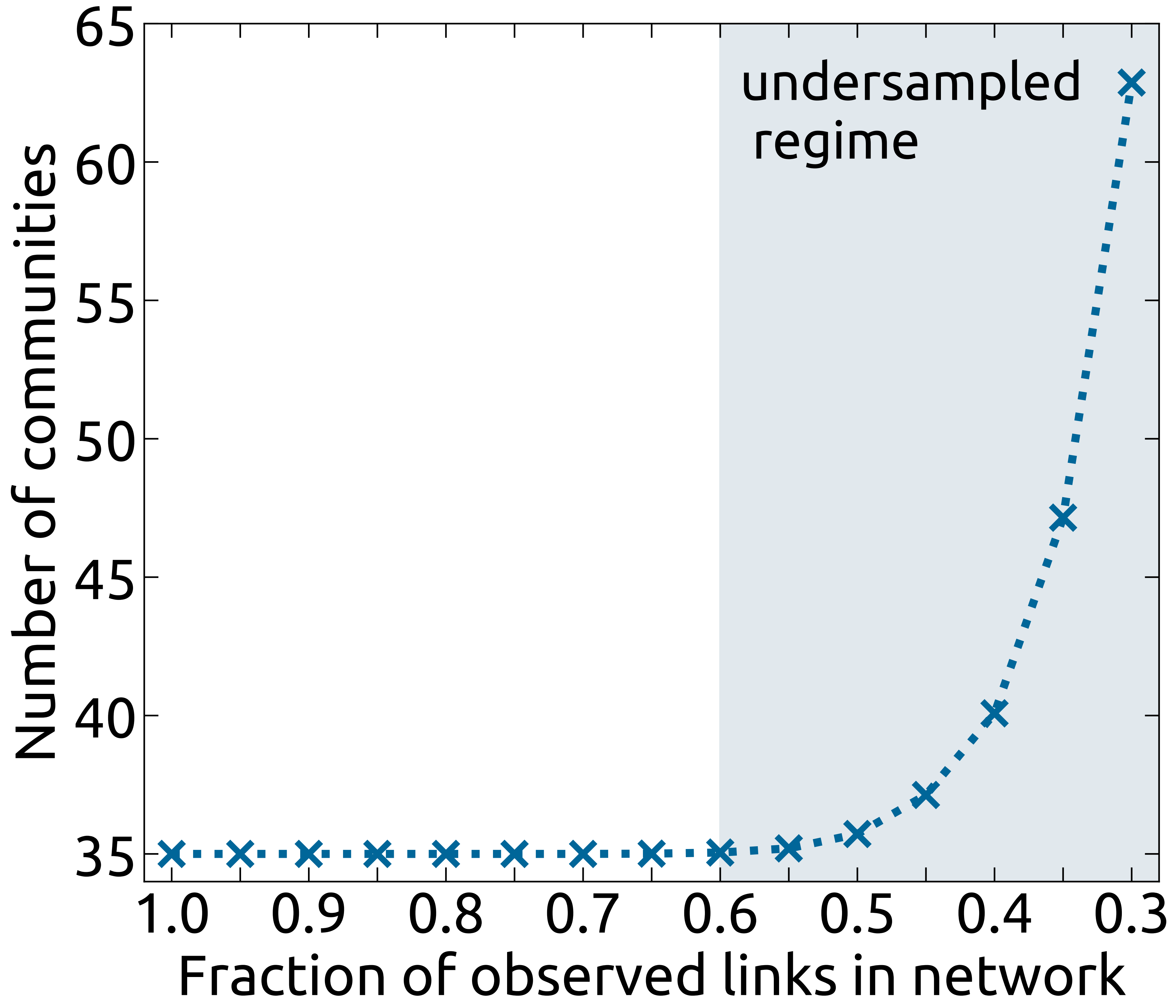
3. Mapping network flows with incomplete information

The map equation

Spurious communities resulting from mere noise



The map equation requires stronger regularization for sparse networks



Bayesian estimate of the map equation

Bayesian estimate of the map equation

$$\hat{L}_B(M) = \int L(M) P(\boldsymbol{\rho} | \text{network data}) d\boldsymbol{\rho}$$

Visit and transition rates

$$\boldsymbol{\rho} = (p_\alpha, q_{i\curvearrowleft}, q_{i\curvearrowright})$$

Posterior probability to observe the rates given the data

$$P(\boldsymbol{\rho} | \text{network data}) = \frac{P(\text{network data} | \boldsymbol{\rho}) P(\boldsymbol{\rho})}{P(\text{network data})}$$

Prior distribution

Prior parameters

$$P(\boldsymbol{\rho}) = P(p_\alpha, q_{i\curvearrowleft}, q_{i\curvearrowright} | a_\alpha, a_{i\curvearrowleft}, a_{i\curvearrowright})$$

Dirichlet distribution

$$= \frac{\Gamma(a_1 + \cdots + a_{m\curvearrowright})}{\Gamma(a_1) \ldots \Gamma(a_{m\curvearrowright})} \prod_{\alpha=1}^{V} p_\alpha^{a_\alpha - 1} \prod_{i=1}^{m} q_{i\curvearrowleft}^{a_{i\curvearrowleft} - 1} \prod_{i=1}^{m} q_{i\curvearrowright}^{a_{i\curvearrowright} - 1}$$

Bayesian estimate of the map equation

Undirected and unweighted networks

$$
\begin{aligned}
\hat{L}_B(M) = \frac{1}{\ln(2)} \frac{1}{\sum_{\alpha=1}^{V} u_\alpha} \times \\
\times [-\sum_{\alpha=1}^{V} u_\alpha \psi(u_\alpha + 1) - \sum_{i=1}^{m} u_{i\curvearrowright} \psi(u_{i\curvearrowright} + 1) \\
+ \sum_{i=1}^{m} (u_{i\curvearrowright} + \sum_{\alpha \in i} u_\alpha) \psi(u_{i\curvearrowright} + \sum_{\alpha \in i} u_\alpha + 1) \\
- \sum_{i=1}^{m} u_{i\curvearrowleft} \psi(u_{i\curvearrowleft} + 1) + (\sum_{i=1}^{m} u_{i\curvearrowleft}) \psi(\sum_{i=1}^{m} u_{i\curvearrowleft} + 1)]
\end{aligned}
$$

$$u_x = k_x + a_x$$

Bayesian estimate of the map equation

Choosing a prior distribution

Prior assumption:

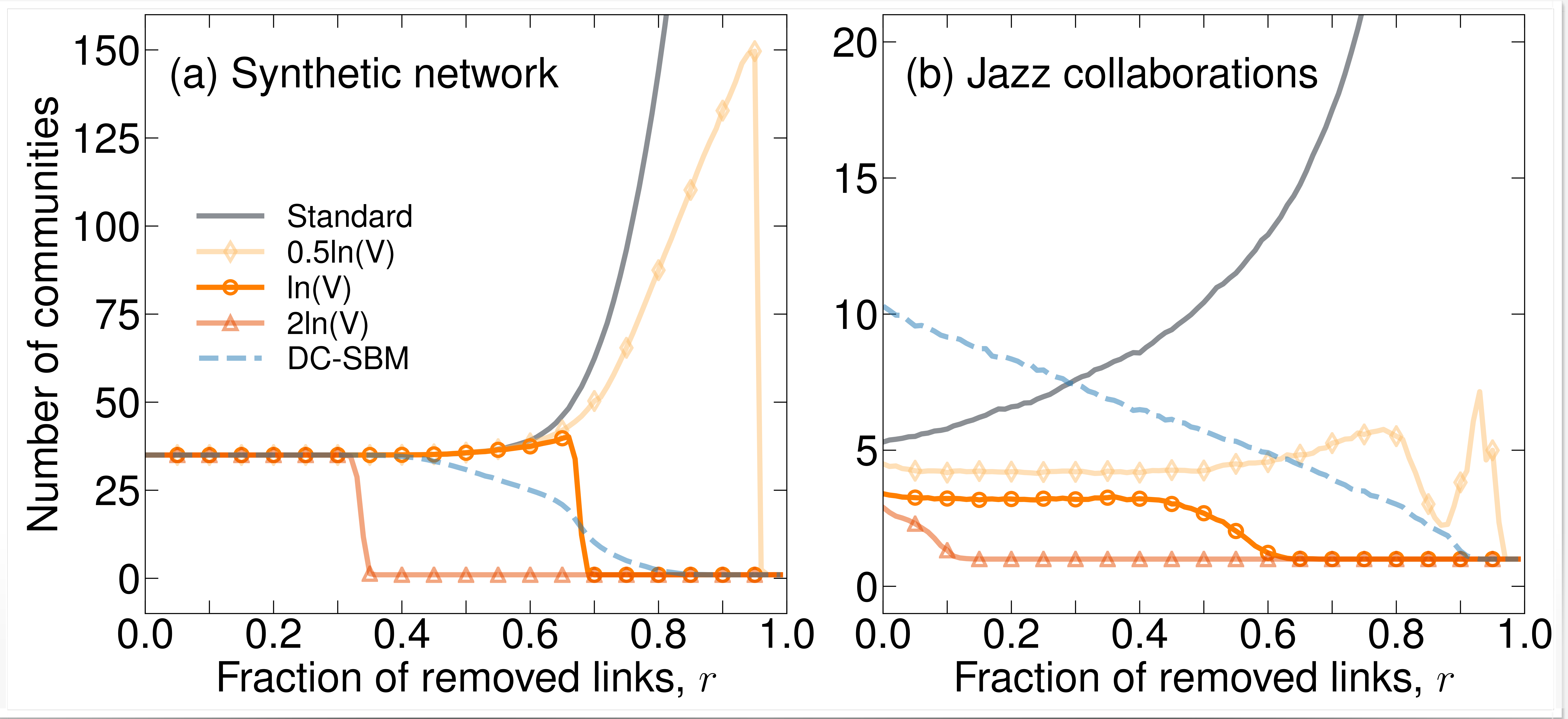
- random network
- each pair of nodes connected with probability $p = \frac{a}{V-1}$

$\Rightarrow$ Parameters of the Dirichlet distribution:

- $a_\alpha = a$
- $a^i_{\curvearrowleft} = a^i_{\curvearrowright} = aV_i \frac{V-V_i}{V-1}$

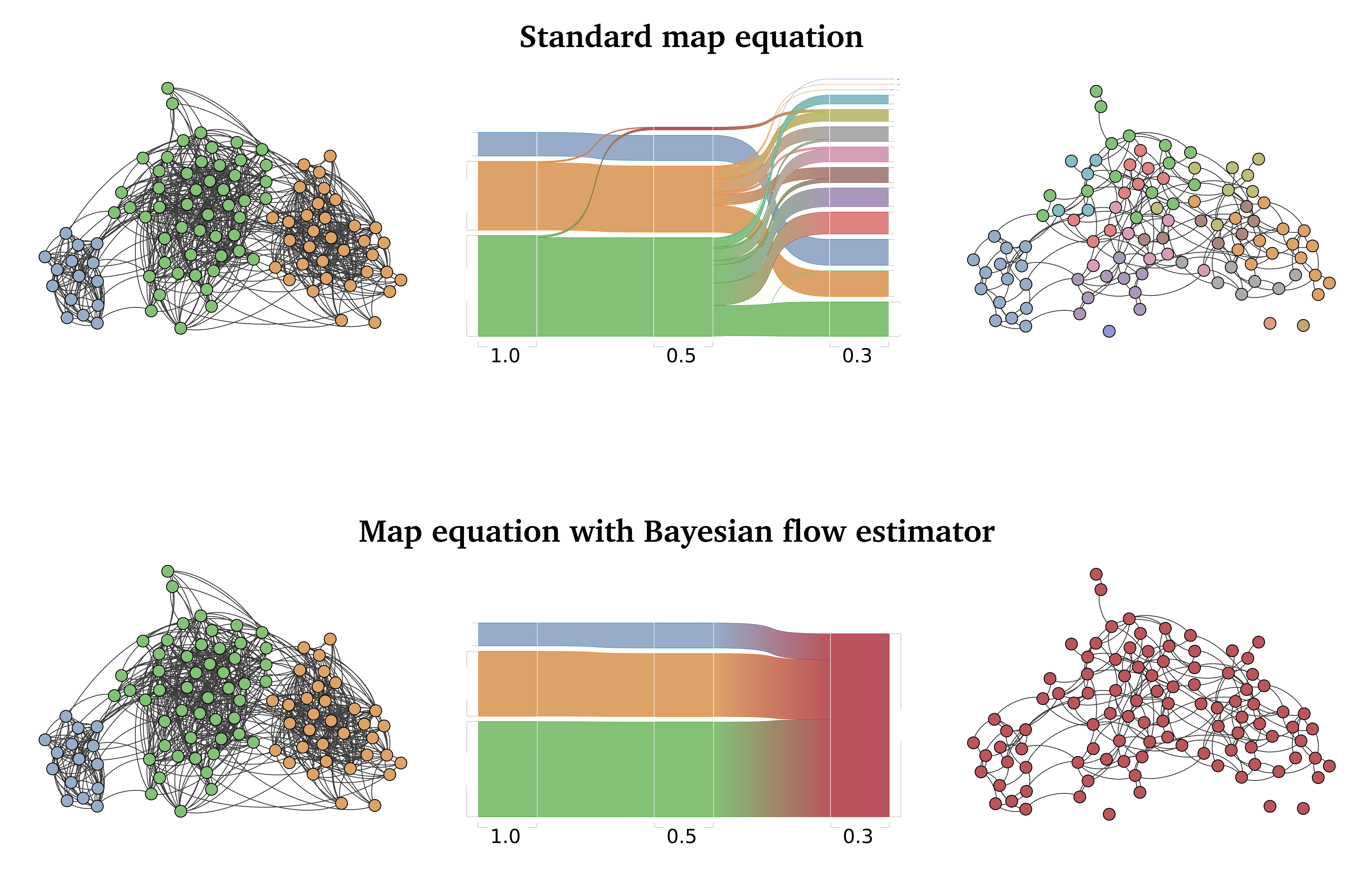
Bayesian estimate of the map equation
Choosing a prior distribution

$$a = C \log V$$



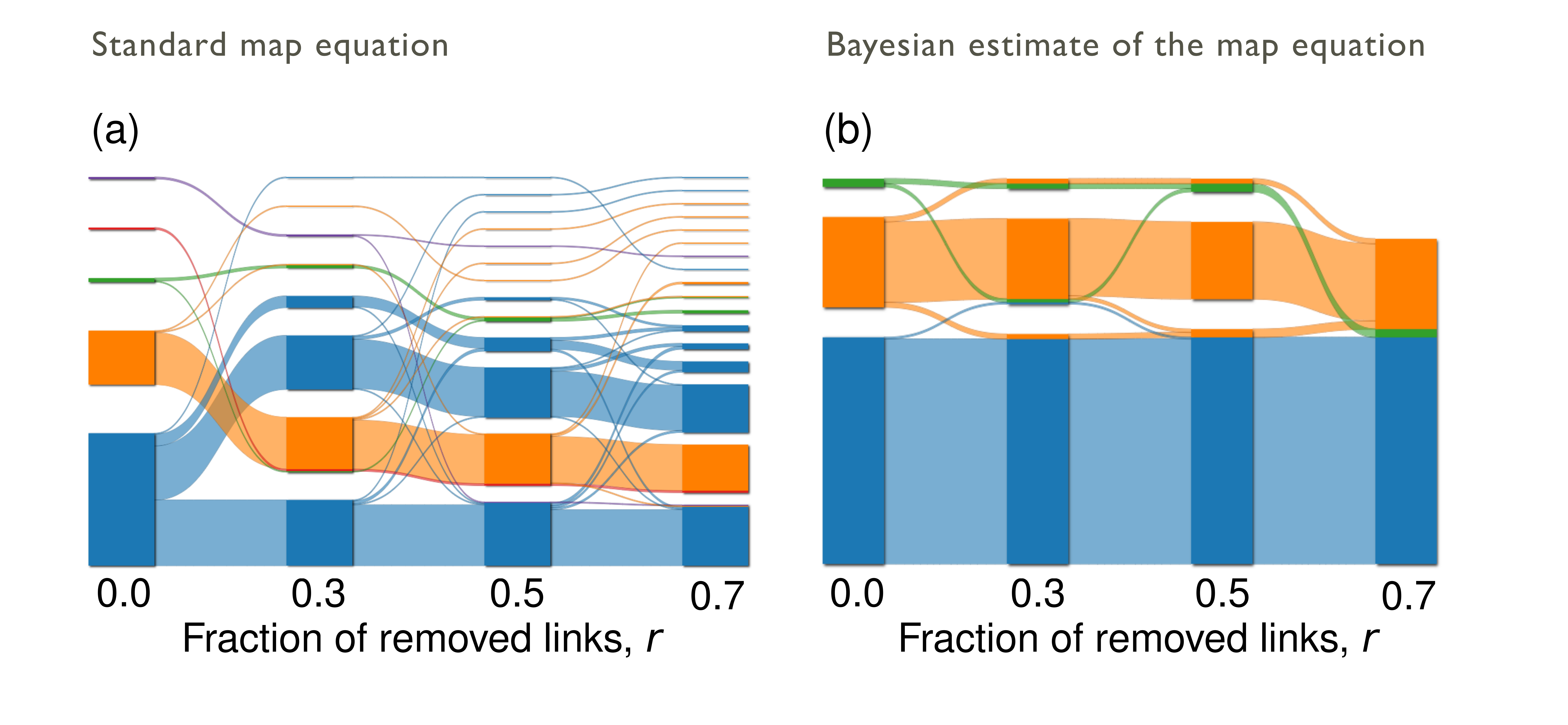
Bayesian estimate of the map equation

Choosing a prior distribution



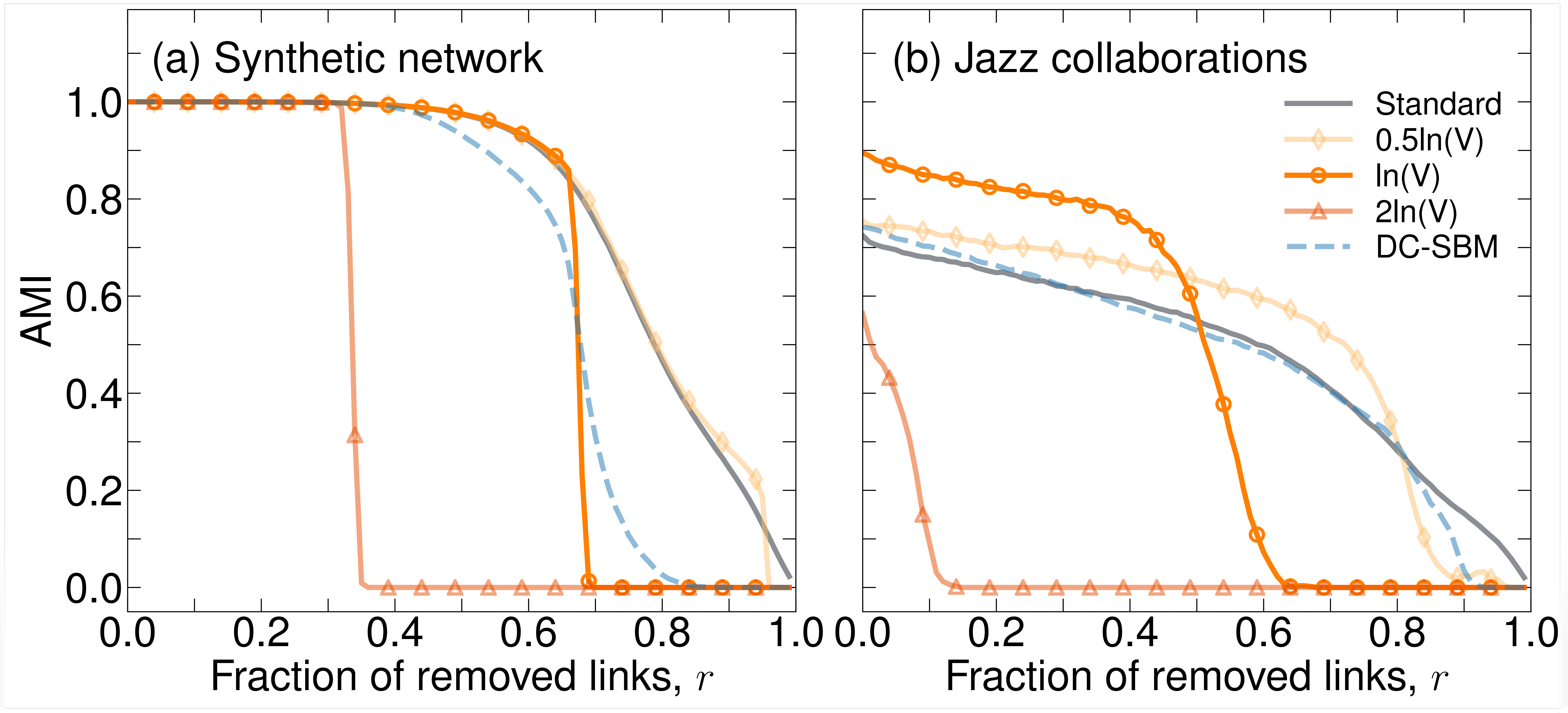
Bayesian estimate of the map equation

Jazz collaboration network



Bayesian estimate of the map equation

Adjusted mutual information



Bayesian estimate of the map equation

Phys Rev E 102, 012302 (2020)

PHYSICAL REVIEW E **102**, 012302 (2020)

Mapping flows on sparse networks with missing links

Jelena Smiljanić,[1,2,*] Daniel Edler,[1,3,4] and Martin Rosvall[1]

[1]*Integrated Science Lab, Department of Physics, Umeå University, SE-901 87 Umeå, Sweden*

[2]*Scientific Computing Laboratory, Center for the Study of Complex Systems, Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia*

[3]*Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Gothenburg, Sweden*

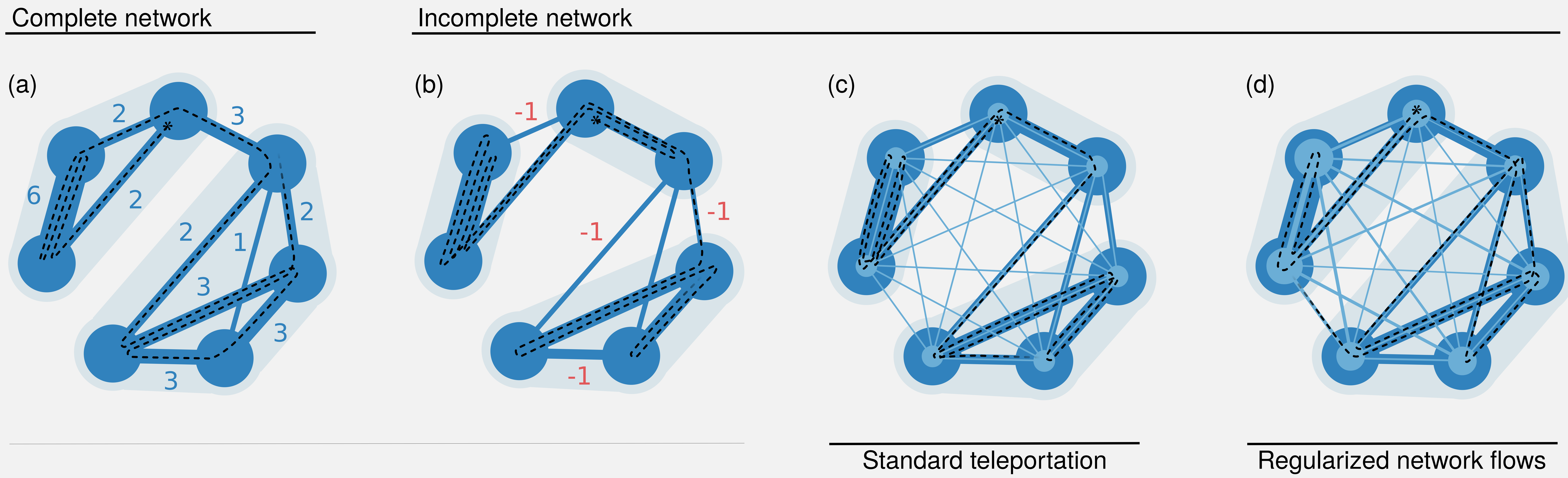
[4]*Department of Biological and Environmental Sciences, University of Gothenburg, Carl Skottsbergs gata 22B, Gothenburg 41319, Sweden*

(Received 12 December 2019; revised 12 May 2020; accepted 9 June 2020; published 6 July 2020)

Unreliable network data can cause community-detection methods to overfit and highlight spurious structures with misleading information about the organization and function of complex systems. Here we show how to detect significant flow-based communities in sparse networks with missing links using the map equation. Since the map equation builds on Shannon entropy estimation, it assumes complete data such that analyzing undersampled networks can lead to overfitting. To overcome this problem, we incorporate a Bayesian approach with assumptions about network uncertainties into the map equation framework. Results in both synthetic and real-world networks show that the Bayesian estimate of the map equation provides a principled approach to revealing significant structures in undersampled networks.

Bayesian take on the map equation

Directed and weighted networks



An empirical Bayes estimate of the transition rates
for the map equation

Bayesian estimate of the transition rates

$$\hat{t}_{ij}(W_i) = \int t_{ij} P(T_i|W_i) dT_i$$

Maximum likelihood estimate

$$\left(\tilde{t}_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}\right)$$

Posterior probability to observe the transition rates given the data

$$P(T_i|W_i) = \frac{P(W_i|T_i)P(T_i)}{P(W_i)}$$

Likelihood of the data given the Dirichlet distribution as prior

$$P(W_i|T_i) = (w_{i1} + \ldots + w_{iN})! \prod_{j=1}^{N} \frac{t_{ij}^{w_{ij}}}{w_{ij}!}$$

Probability of the data

$$P(W_i) = \int P(W_i|T_i)P(T_i) dT_i$$

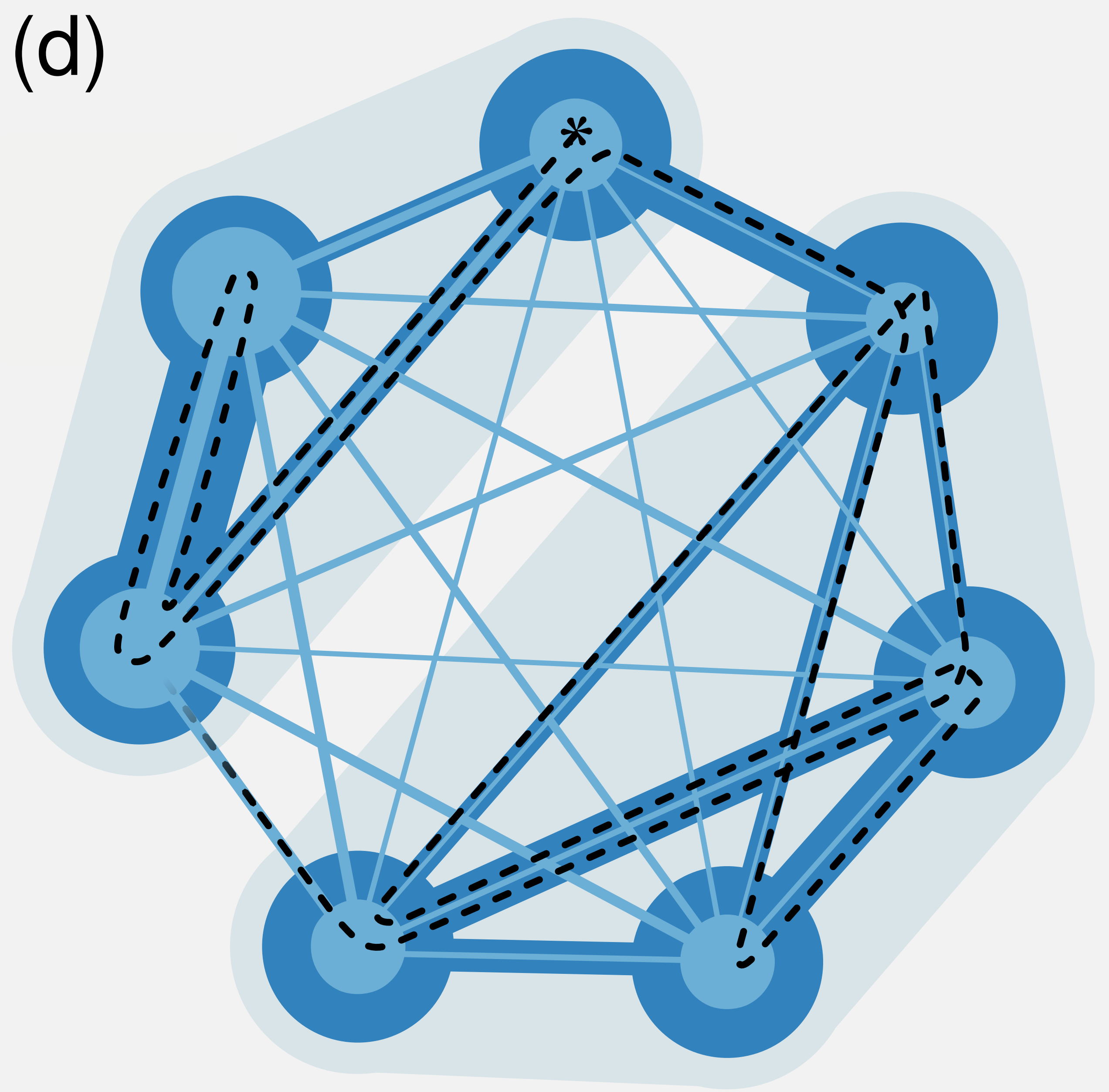
Posterior distribution

$$P(T_i|W_i, \gamma_i) \propto \prod_{j=1}^{N} t_{ij}^{w_{ij}+\gamma_{ij}-1}$$

Posterior distribution after integrating

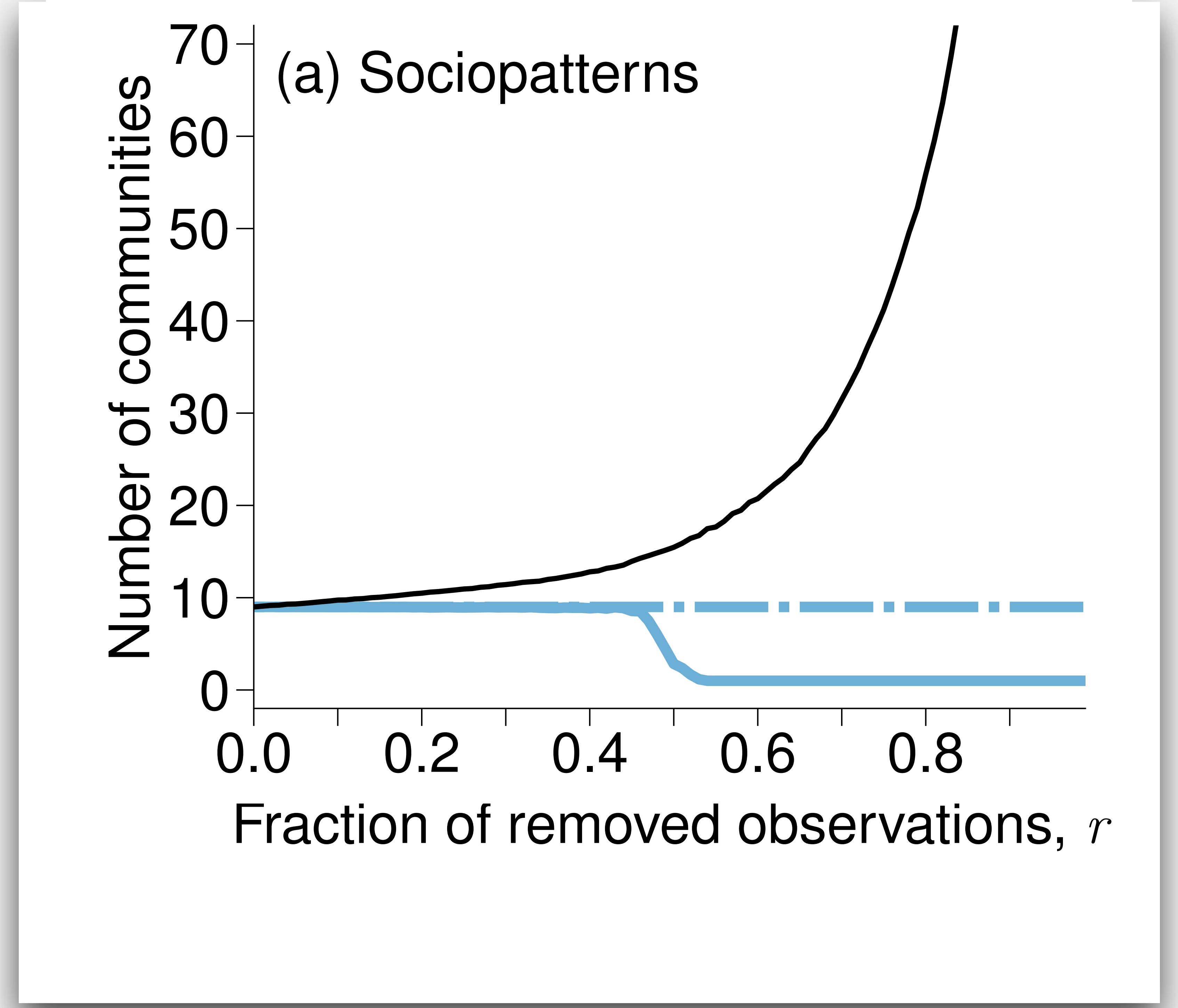
$$(1-\alpha_i)\frac{w_{ij}}{\sum_j w_{ij}} + \alpha_i \frac{\gamma_{ij}}{\sum_j \gamma_{ij}} \text{ for } \alpha_i = \frac{\sum_{j=1}^{N} \gamma_{ij}}{\sum_{j=1}^{N} w_{ij} + \gamma_{ij}}$$

$$\gamma_{ij} = \frac{\ln N}{N} \frac{\sum_{n=1}^{N} k_n^{\text{in}} + k_n^{\text{out}}}{\sum_{n=1}^{N} s_n^{\text{in}} + s_n^{\text{out}}} \frac{s_i^{\text{out}} s_j^{\text{in}}}{k_i^{\text{out}} k_j^{\text{in}}}$$



Regularized network flows

An empirical Bayes estimate of the transition rates
for the map equation



Recorded interactions between female and male students in a high school in Marseille.

The students are assigned to one of 9 classes, which we use as metadata.

An empirical Bayes estimate of the transition rates
for the map equation

Mapping flows on weighted and directed networks with incomplete observations

Jelena Smiljanić,[1,2,*] Christopher Blöcker,[1] Daniel Edler,[1,3,4] and Martin Rosvall[1]
[1]*Integrated Science Lab, Department of Physics, Umeå University, SE-901 87 Umeå, Sweden*
[2]*Scientific Computing Laboratory, Center for the Study of Complex Systems, Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia.*
[3]*Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Gothenburg, Sweden.*
[4]*Department of Biological and Environmental Sciences, University of Gothenburg, Carl Skottsbergs gata 22B, Gothenburg 41319, Sweden.*
(Dated: November 2, 2021)

Detecting significant community structure in networks with incomplete observations is challenging because the evidence for specific solutions fades away with missing data. For example, recent research shows that flow-based community detection methods can highlight spurious communities in sparse undirected and unweighted networks with missing links. Current Bayesian approaches developed to overcome this problem do not work for incomplete observations in weighted and directed networks that describe network flows. To address this gap, we extend the idea behind the Bayesian estimate of the map equation for unweighted and undirected networks to enable more robust community detection in weighted and directed networks. We derive a weighted and directed prior network that can incorporate metadata information and show how an efficient implementation in the community-detection method Infomap provides more reliable communities even with a significant fraction of data missing.

I. INTRODUCTION

tection in directed and weighted networks remains unresolved.

3. Mapping network flows with incomplete information

The Bayesian estimate of the map equation provides a principled approach to revealing significant structures in undersampled networks.

Thank you!

www.mapequation.org/infomap/