Mapping nonlocal relationships between metadata and network structure with metadata-dependent encoding of random walks

Martin Rosvall
ELLlIT Focus Period 2023 on Network Dynamics and Control
Mapping network structure
Mapping network structure

**Function 1**
**Exploration**

- Geographic Maps
- Google Maps for networks

**Simplify and highlight**
- Organization
- Function
- Understanding

- Network theory + information theory

**Function 2**
**Navigation**

- Show directions
- Recommendations
- Drug repurposing
- Precision medicine

- Network theory + information theory + machine learning

---

**MapEq**

- Modular code books
- Modular path costs in bits

---

**MapSim**

- Modular code books
- Modular path costs in bits
The map equation framework with a Bayesian estimate of the transition rates provides a principled approach to mapping flows on multilayer networks with incomplete observations.
Mapping network structure and metadata on power grids

“We would like to play with some parameters to give more importance to some nodes in the graph, somehow conditioning the clustering when we are interested in specific elements of the graph.

Antoine Marot
Lead AI Scientist at RTE – France’s Transmission System Operator
Mapping **network structure and metadata** on power grids

“We want coherent communities with nodes that share similar prices.”

Antoine Marot
Lead AI Scientist at RTE – France’s Transmission System Operator
**Question:** How can we exploit nonlocal relationships between network structure and metadata?

---

**1. INTRODUCTION**

As network science has found application in a variety of real-world systems, ranging from the biological to the technological, so, too, has community detection in networks received widespread attention [1–4]. Traditionally, community detection methods have focused solely on the topology of the network, optimizing an objective function defined on the network structure that captures a particular notion of community, such as intracommunity edge density and intercommunity edge sparsity. Many approaches, ranging from the statistical to the information theoretical, have been used for community detection. From the point of view of this framework, our method’s contribution is enabling users to choose how much control the metadata’s relative importance for identifying network structure. On synthetic networks, we show that our algorithm can overcome the structural detectability limit when the metadata are well aligned with community structure.

On real-world networks, we show how our algorithm can achieve greater mutual information with the network structure than can be done with either alone. Crucially, the method does not assume that the metadata are correlated with the network and its metadata to detect communities more accurately than can be done with either alone. Instead, the method learns whether a correlation exists either alone. Our tuning parameter allows users to “zoom in” and “zoom out” on communities with varying levels of focus on the metadata. One of the well-known Louvain algorithm [5] received widespread attention [6–8]. Similarly, Hric et al. [9] developed a directed stochastic block model with multiple continuous attributes. Stanley et al. [10] proposed CNMMA, which weights communities aligned with gender to communities aligned with a particular metadata type. With our method, the metadata at a cost in the traditional map equation. Our tuning parameter, like the focusing knob of a microscope, allows users to “zoom in” and “zoom out” on communities with varying levels of focus on the metadata.
1. Mapping network flows

2. with metadata-dependent encoding

3. exploits nonlocal relationships
Mapping network flows using the minimum description length principle
MAPS depict regularities using less information.

NETWORKS describe where flows move depending on the current node.
Mapping network flows using the minimum description length principle

Fig. 1. — Schematic diagram of a general communication system.
Mapping network flows using the minimum description length principle.
Mapping network flows
using the minimum description length principle

$L(M_1) = H(\mathcal{P}) = 4.23$ bits

$L(M_5) = q \cdot H(\mathcal{Q}) + \sum_{i=1}^{5} p_i \cdot H(\mathcal{P}_i) = 0.42 + 3.13 = 3.55$ bits
Mapping network flows using the map equation

Visit and transition rates
\[ \rho = (p_\alpha, q_i, q_i^\cap) \]

Per-step use rate of the index codebook.
\[ q^\cap = \sum_{i=1}^{m} q_i \]

Per-step use rate of module codebook \( i \)
\[ p_i^\cup = q_i^\cap + \sum_{\alpha \in M_i} p_\alpha \]

Per-step average code length of index codebook for steps between modules
\[ L(M) = q^\cap H(Q) + \sum_{i=1}^{m} p_i^\cup H(P_i) \]

Where
- \( Q = \{q^\cap\} \)
- \( P_i = \{q_i^\cap, p_\alpha \in M_i\} \)
The map equation
Compression of network flows

\[ L(M) = H(P) = 4.75 \text{ bits.} \]

\[ L(M) = q_{\rightarrow} H(Q) + p_{\rightarrow} H(P^1) + p_{\rightarrow} H(P^2) + p_{\rightarrow} H(P^3) = 3.68 \text{ bits.} \]

\[ L(M) = q_{\rightarrow} H(Q) + p_{\rightarrow} H(P^1) + p_{\rightarrow} H(P^2) + p_{\rightarrow} H(P^3) = 3.57 \text{ bits.} \]
The map equation

Compression of network flows

1

\( L(M) = q_X H(Q) + \{ p_1^0 H(P^1) \\ p_1^2 H(P^2) \\ p_1^3 H(P^3) \} = 3.68 \text{ bits.} \)

22

\( L(M) = q_X H(Q) + \{ p_5^0 H(P^1) \\ p_5^2 H(P^2) \\ p_5^3 H(P^3) \} = 3.57 \text{ bits.} \)

3

\( L(M) = q_X H(Q) + \{ p_6^0 H(P^1) \\ p_6^2 H(P^2) \\ p_6^3 H(P^3) \} + 3.56 \text{ bits.} \)

4

\( q_3^1 H(Q^1) + \{ p_7^0 H(P^{11}) \\ p_7^2 H(P^{12}) \\ p_7^3 H(P^{13}) \} = 3.48 \text{ bits.} \)

5

\( q_5^2 H(Q^2) + \{ p_8^0 H(P^{21}) \\ p_8^2 H(P^{22}) \\ p_8^3 H(P^{23}) \} = 3.48 \text{ bits.} \)

6

\( q_6^3 H(Q^3) + \{ p_9^0 H(P^{31}) \\ p_9^2 H(P^{32}) \\ p_9^3 H(P^{33}) \} = 3.48 \text{ bits.} \)

26

\( 0.12 \text{ bits} \)

\( 0.97 \text{ bits} \)

\( 0.12 \text{ bits} \)

\( 0.76 \text{ bits} \)

\( 2.60 \text{ bits} \)

\( 2.60 \text{ bits} \)
Mapping network flows using the minimum description length principle

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

Generalizations to many network representations.
2. Mapping network flows with metadata-dependent encoding
1. We use random walks that remember their origin

2. Each node $i$ has associated metadata $f_i$

3. The probability $x_{ij}$ of a walker starting at $i$ to be absorbed at $j$ depends on $f_i$ and $f_j$
Mapping network flows

Encoding probabilities for categorical metadata

\[ \varepsilon_{ij} = p \delta_{f_i, f_j} + \frac{p}{c} (1 - \delta_{f_i, f_j}) \]

- If \( c > 1 \), the walker will encode more frequently at nodes belonging to the same class of the starting node (assortative encoding).
- For \( p < c < 1 \), encode will be more probable at nodes belonging to a different class than the one of the starting node (disassortative encoding).
- For \( c = 1 \), the encoding dynamics no longer depend on class assignments (neutral encoding).
- When \( p \ll 1 \), the structure is irrelevant and absorption depends only on metadata.
Mapping network flows

Random walks with metadata-dependent encoding probabilities

We consider a connected and possibly weighted graph. The random walks are colored by the currently encoded module and labeled with circles when they encode a trajectory to a node as a function of its metadata and the previously encoded node, leading to fewer, larger communities with larger diameter. To explore relationships between network structure and metadata, we encode the random walker's transitions of the random walk are encoded. For Markov times larger than 1, the random walk can make more than one transition to a node as a function of its metadata and the previously encoded node, leading to fewer, larger communities with stability. To remain trapped for relatively long times in densely connected subgraphs, we take advantage of a multiscale extension of the map equation that modifies the standard Markov process with one-step encoding probabilities. We interpolate between these two extremes by considering a random graph, then the statistics of the symbolic dynamics sum to one and the metadata-informed communities contain nodes that appear in relatively long uninterrupted encoding sequences (Fig. 1).

In the absence of metadata, the symbolic dynamics are trivial. However, the underlying random walk statistics depend on the structure of the underlying distribution of metadata at different scales. Conversely, if the graph is undirected, but a similar reasoning holds for primitive structural network information or metadata clustering alone. For example, analyzing the spatial network of energy prices across Europe and categorical, scalar, or vectorial metadata, such as gender, occupation, economic similarities.

In this schematic example with single, long random walks, we encode the random walks' next transition and with dotted lines when the metadata differ from the metadata of the previously encoded node. The corresponding metadata-dependent encodings in a metadata-dependent encoding graph where all probabilities of the underlying distribution of metadata at different scales are interpolated between these two extremes by considering a random graph, then the statistics of the symbolic dynamics sum to one and the metadata-informed communities contain nodes that appear in relatively long uninterrupted encoding sequences (Fig. 1).

Constructing the metadata-dependent encoding graph from a random graph, then the statistics of the symbolic dynamics sum to one and the metadata-informed communities contain nodes that appear in relatively long uninterrupted encoding sequences (Fig. 1).

The alphabetic codes represent the walks with this metadata-dependent encoding scheme. The map equation encodes the random walk more efficiently than the solution in (3, 10) with c ∈ (3, 10) and less than 10% gives the solution in (10, 33%) and between 10 and 33% gives the solution in (33, 100%) and greater than 100% gives the solution in (100, 100%).

For simplicity, we assume that edges. For example, analyzing the spatial network of energy prices across Europe and categorical, scalar, or vectorial metadata, such as gender, occupation, economic similarities.
Random walks with metadata-dependent encoding probabilities

Mapping network flows

Original graph $G ightarrow E$ with $E_i = \sum_{t=1}^{\infty} E_i \Pi_t P(t|i)$,

Encoding probabilities

Stationary occupation probability

Transition matrix

Encoding graph
Mapping network flows
Random walks with metadata-dependent encoding probabilities

Synthetic example

A  Original graph

B  $\rho = 1 \quad c = 1$

C  $\rho = 0.5 \quad c = 50$

D  $\rho = 0.1 \quad c = 100$
Encoding probabilities for real-valued metadata

\[ \varepsilon_{ij} = s \exp \left( \frac{-|f_i - f_j|}{b} \right) p + (1 - s) \]

- Encoding probability
- Real-valued node metadata
- Baseline encoding probability
- Strength of metadata information
- Scale parameter
Random walks with metadata-dependent encoding probabilities integrate network information and distant metadata.
Mapping network flows with metadata-dependent encoding exploits nonlocal relationships.
Starting point matters
Income class matters in the commuting network of London

(a) \( r_m = 0.22 \), \( p/c = 0.8 \)
(b) \( r_m = 0.23 \), \( p/c = 0.6 \)
(c) \( r_m = 0.88 \), \( p/c = 0.25 \)
(d) \( r_m = 1.0 \), \( p/c = 0.1 \)

Class
Unemployment class matters in the commuting network of London

Unemployment class matters in the commuting network of London.
Obesity class matters in the commuting network of London

![Maps showing different obesities](https://example.com/maps)

- **A** $r_m = 0.26$, $p/c = 0.8$
- **B** $r_m = 0.3$, $p/c = 0.6$
- **C** $r_m = 0.57$, $p/c = 0.25$
- **D** $r_m = 1.0$, $p/c = 0.1$

**Class**

Downloaded from https://www.science.org on September 20, 2023
European power grid network with node prices and optimal partitions

Panel A: Node prices (€)

Panel B: $s = 0$

Panel C: $s = 0.3$

Panel D: $s = 1$
Mapping nonlocal relationships between metadata and network structure with metadata-dependent encoding of random walks

Aleix Bassolas$^{1,2,3}$, Anton Holmgren$^4$, Antoine Marot$^5$, Martin Rosvall$^6$, Vincenzo Nicosia$^1$

Integrating structural information and metadata, such as gender, social status, or interests, enriches networks and enables a better understanding of the large-scale structure of complex systems. However, existing approaches to augment networks with metadata for community detection only consider immediately adjacent nodes and cannot exploit the nonlocal relationships between metadata and large-scale network structure present in many spatial and social systems. Here, we develop a flow-based community detection framework based on the map equation that integrates network information and metadata of distant nodes and reveals more complex relationships. We analyze social and spatial networks and find that our methodology can detect functional metadata-informed communities distinct from those derived solely from network information or metadata. For example, in a mobility network of London, we identify communities that reflect the heterogeneity of income distribution, and in a European power grid network, we identify communities that capture relationships between geography and energy prices beyond country borders.

INTRODUCTION

The network structure of a complex system provides meaningful insights into its function, dynamics, and evolution (1–3). For example, partitioning networks into internally densely connected communities or modules of nodes helps researchers understand how systems organize at different scales (4). However, when nodes with similar metadata are far apart in the network structure with and network structure exist, the presence of metadata adds no value to the extended stochastic block models (16, 17). Similarly, encoding metadata in flow-based modules without local correlations diminishes its usefulness. Thus, the art of finding important mesoscale structures in networks is to integrate metadata in community detection, the encoding of link-related metadata with multiple random walks, and to improve the quality and resolution of the two, because they cannot exploit nonlocal relationships between structure and metadata, our framework merges distant nodes and network structure exist, the presence of metadata adds no value to the extended stochastic block models (16, 17). Similarly, encoding metadata in flow-based modules without local correlations diminishes its usefulness. Thus, the art of finding important mesoscale structures in networks is to integrate metadata in community detection, the encoding of link-related metadata with multiple random walks, and to improve the quality and resolution of the mapping of nonlocal relationships between metadata and network structure.
3. Mapping network flows with metadata-dependent encoding exploits nonlocal relationships.

Random walks with metadata-dependent encoding probabilities reveals functional metadata-informed communities.
CONCLUSION

Random walks with metadata-dependent encoding probabilities integrate structural and metadata information beyond nodes’ immediate neighbors, revealing functional metadata-informed communities.
The map equation framework with a Bayesian estimate of the transition rates provides a principled approach to mapping flows on multilayer networks with incomplete observations.
Three ways to run Infomap
https://www.mapequation.org/infomap/

Python  (C++ speed)

To install, run

```
pip install infomap
```

To upgrade, run

```
pip install --upgrade infomap
```

Infomap only supports Python 3
We currently build packages for Python 3.6 to 3.10.

C++

```
git clone git@github.com:mapequation/infomap.git
cd infomap
make -j
```

Infomap online  (JS speed)