Mapping flows in bipartite networks

The bipartite map equation
NetSciX 2020, Tokyo, Japan

Christopher Blöcker and Martin Rosvall
Integrated Science Lab (IceLab)
Department of Physics
Umeå University, Sweden

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

UMEÅ UNIVERSITY

icelab INTEGRATED SCIENCE LAB

The map equation

- Information-theoretic objective function
- Exploits duality between compression and finding patterns
- Models information flows with a random walk

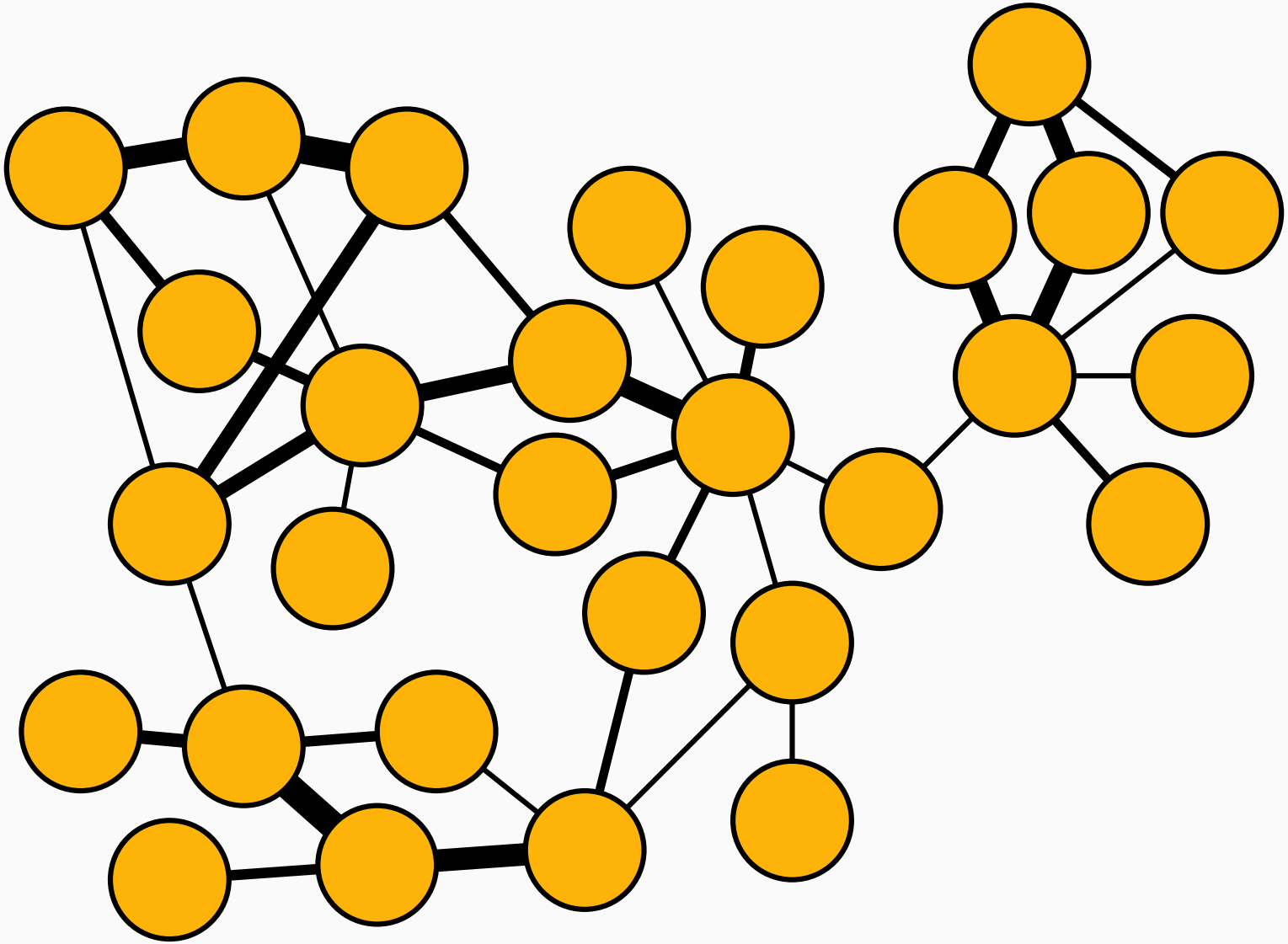
Communication game

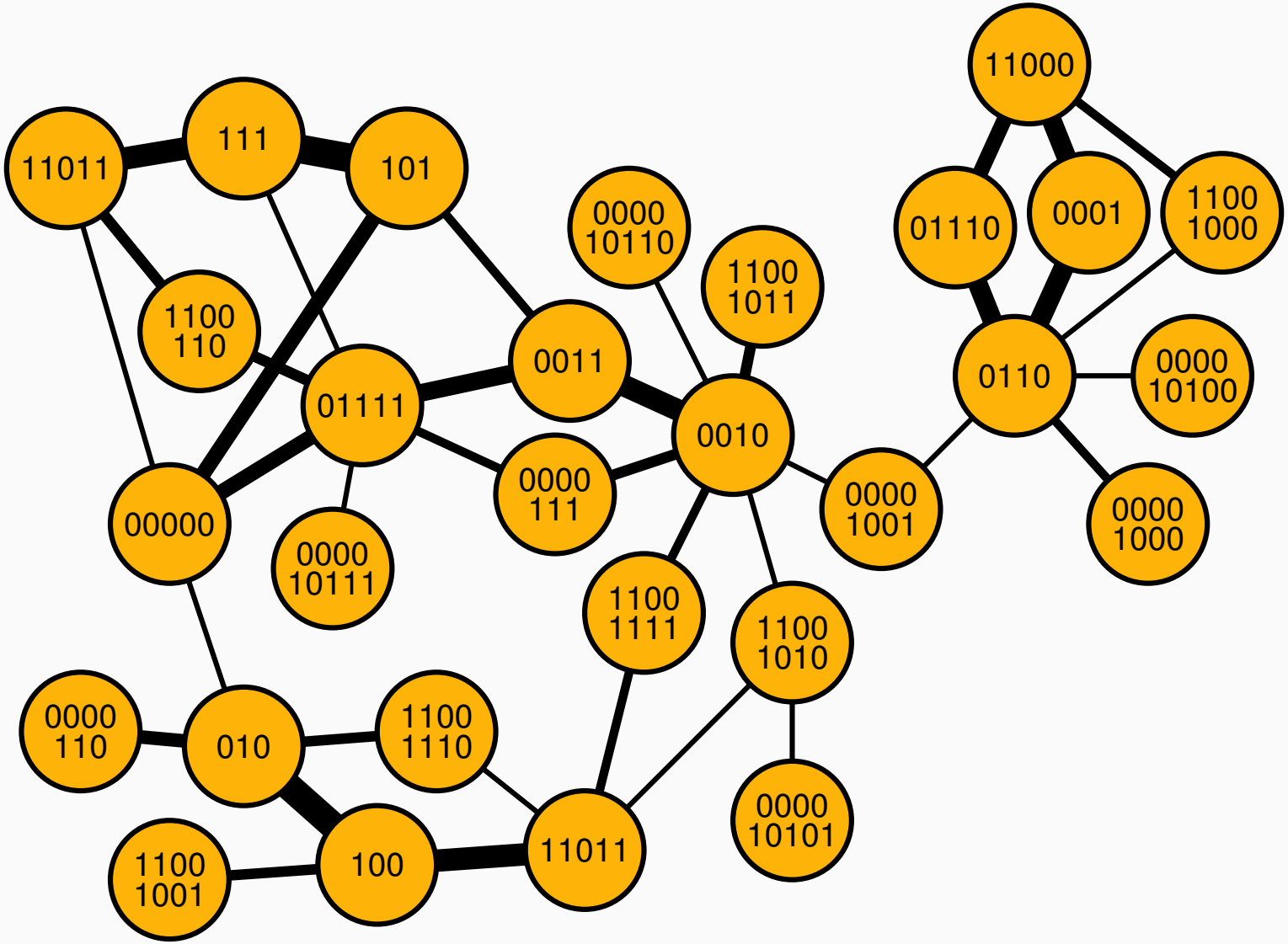
- Sender
- Receiver
- Random walker in a network
- Sender updates receiver about random walker location

Question

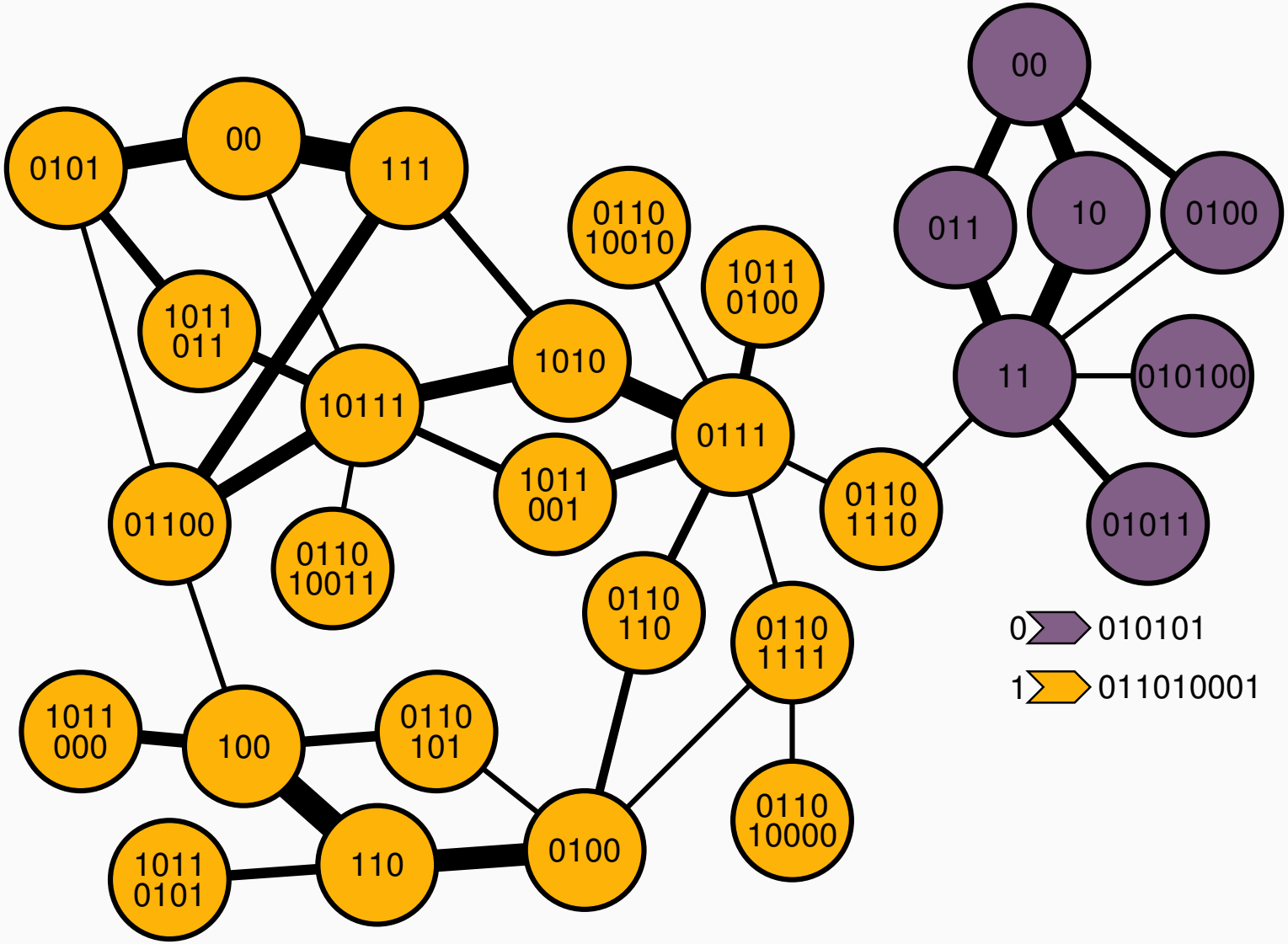
How can we design an efficient coding scheme for this?

efficient = minimise $\underbrace{\text{average per-step description length}}_{\text{“code length”}}$

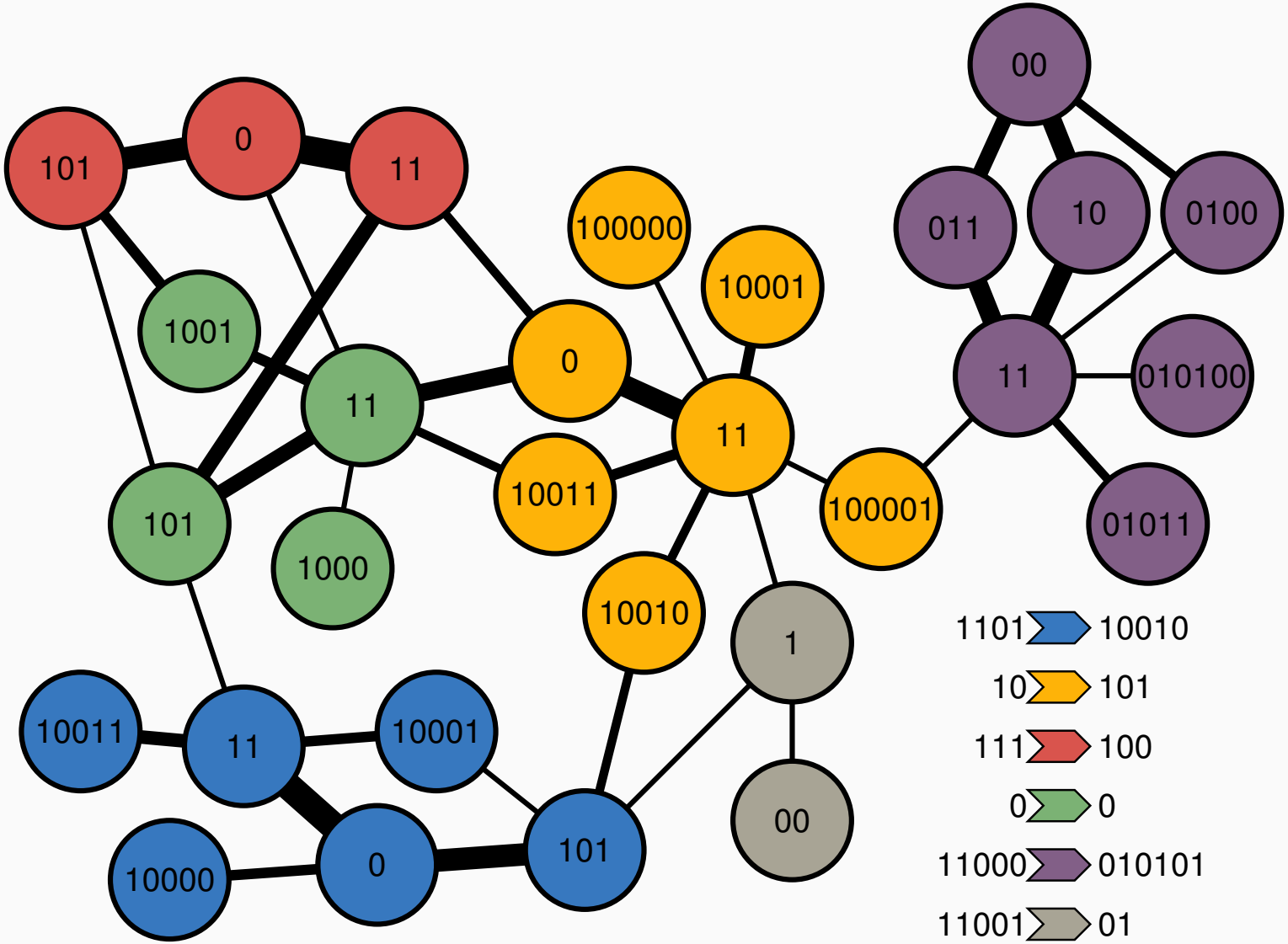




What is the code length? $L(\mathsf{M}) = H(\mathcal{P}) = 3.85\text{bits}$



$$L(\mathsf{M}) = qH(\mathcal{Q}) + p_1 H(\mathcal{P}_1) + p_2 H(\mathcal{P}_2) = 3.22\text{bits}$$



$$L(\mathsf{M}) = qH(\mathcal{Q}) + \sum_{i=1}^{m} p_i H(\mathcal{P}_i) = 2.24\text{bits}$$

The map equation

$$L(\mathsf{M}) = qH(\mathcal{Q}) + \sum_{i=1}^{m} p_i H(\mathcal{P}_i)$$

...can be generalised through recursion

What is the problem?

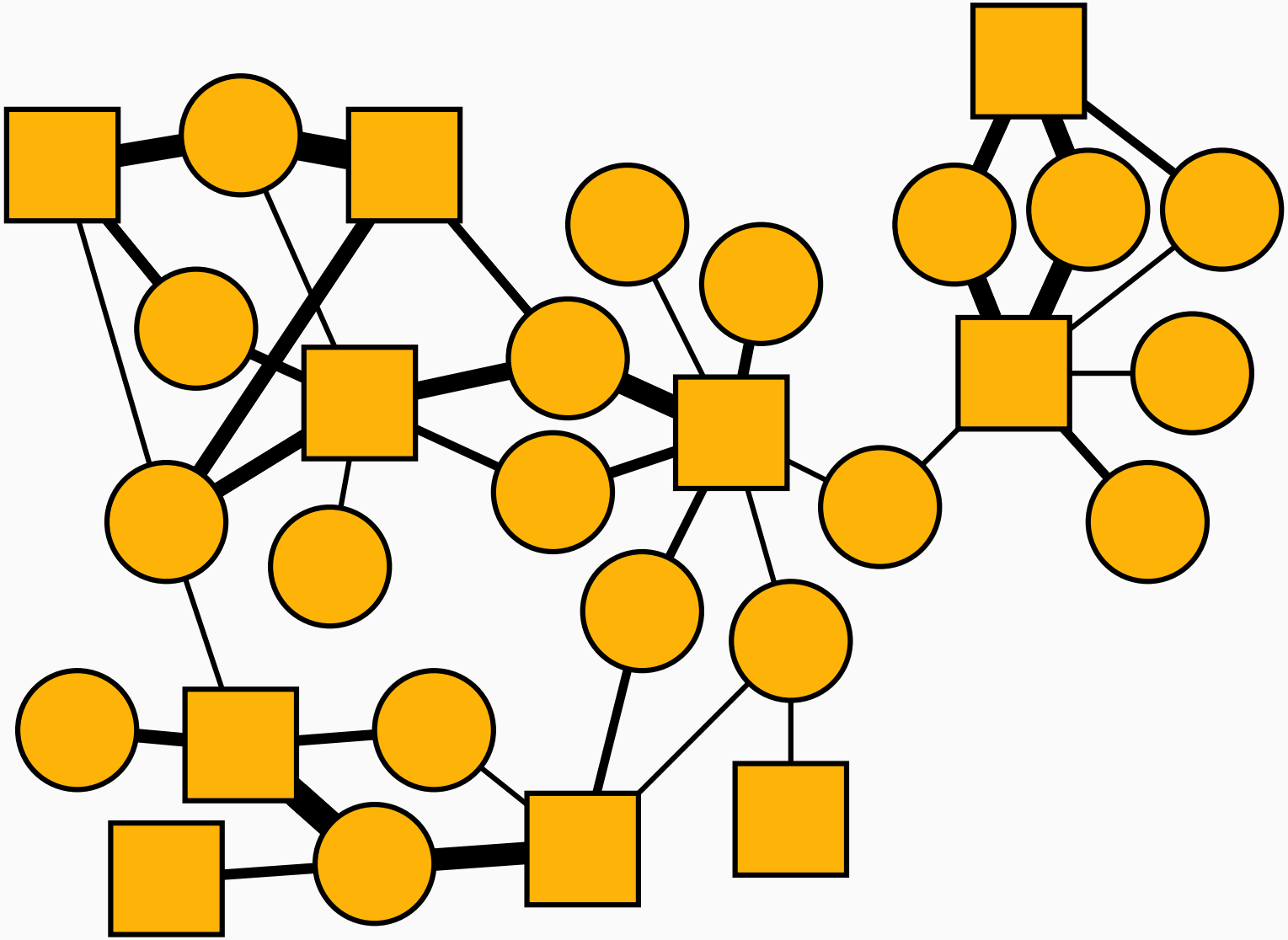
- Coding scheme for steps between any pair of nodes
- Random walks in **bipartite** networks **alternate** between node types, this excludes many transitions

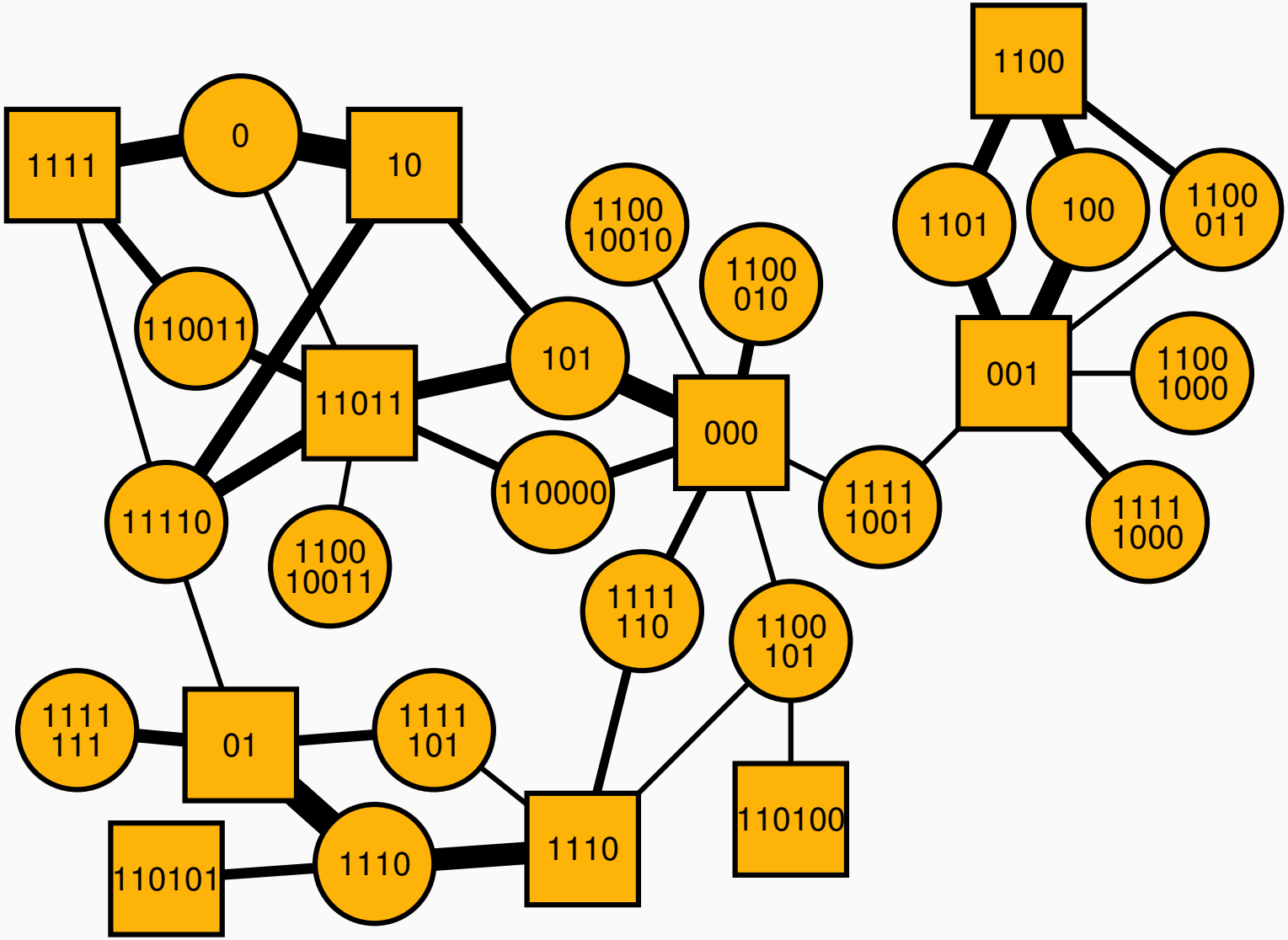
→ The map equation produces longer codes than necessary

Does that mean we cannot use it on bipartite networks?

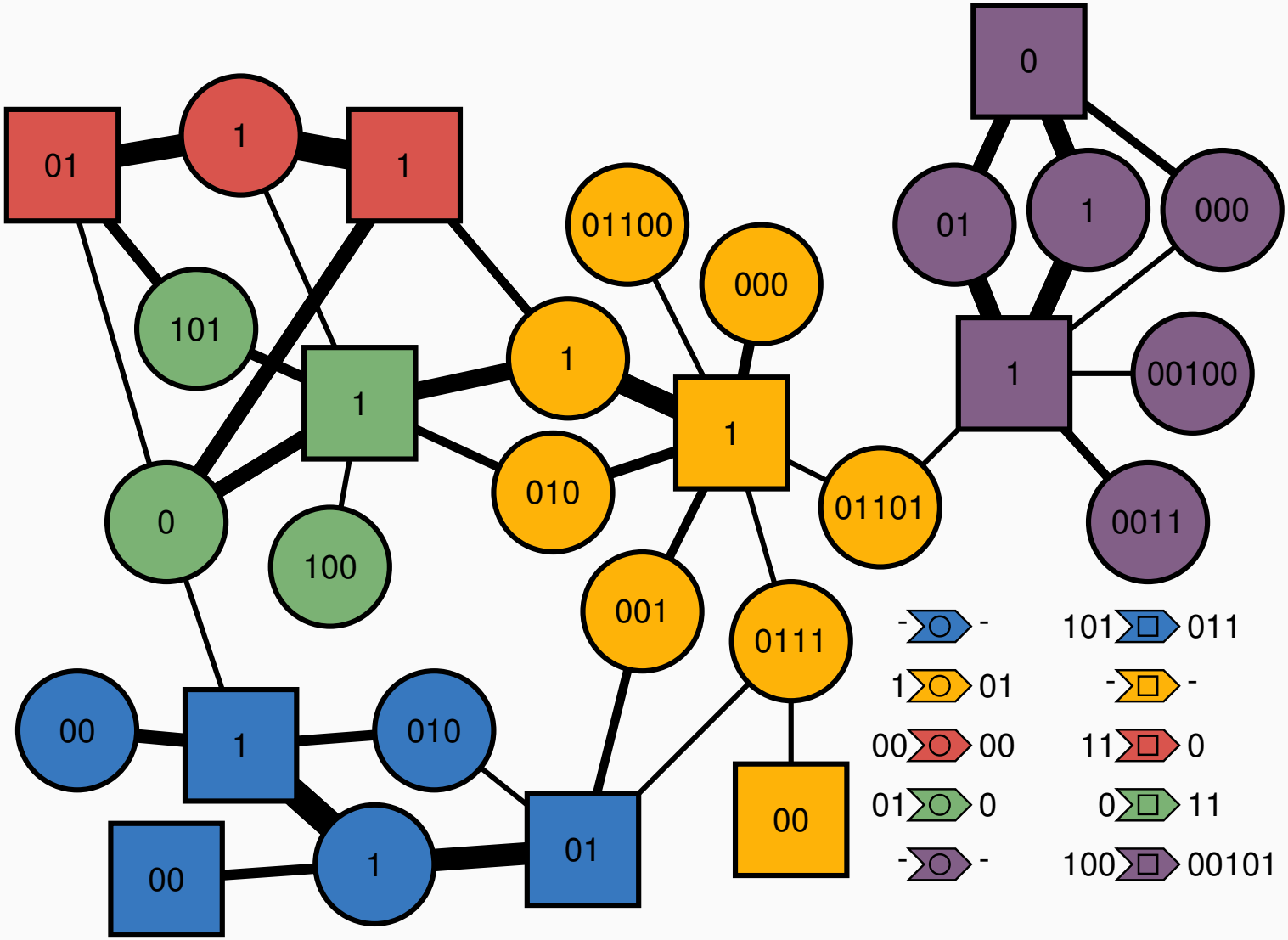
- We can make a unipartite projection, but that's lossy...
- We can use the map equation directly, but that's wasteful...

→ How can we do better?





$$L(\mathsf{M}) = \frac{1}{2}H\left(\mathcal{P}^1\right) + \frac{1}{2}H\left(\mathcal{P}^2\right) = 2.85\text{bits}$$



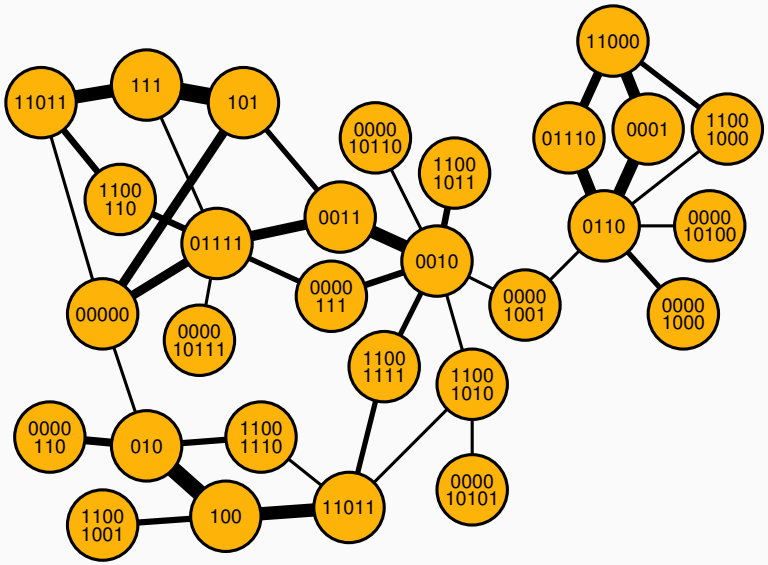
$$L_B(\mathsf{M}) = q^1 H\left(\mathcal{Q}^1\right) + \sum_{i=1}^{m} p_i^1 H\left(\mathcal{P}_i^1\right) + q^2 H\left(\mathcal{Q}^2\right) + \sum_{i=1}^{m} p_i^2 H\left(\mathcal{P}_i^2\right)$$

$$= 1.18\text{bits}$$

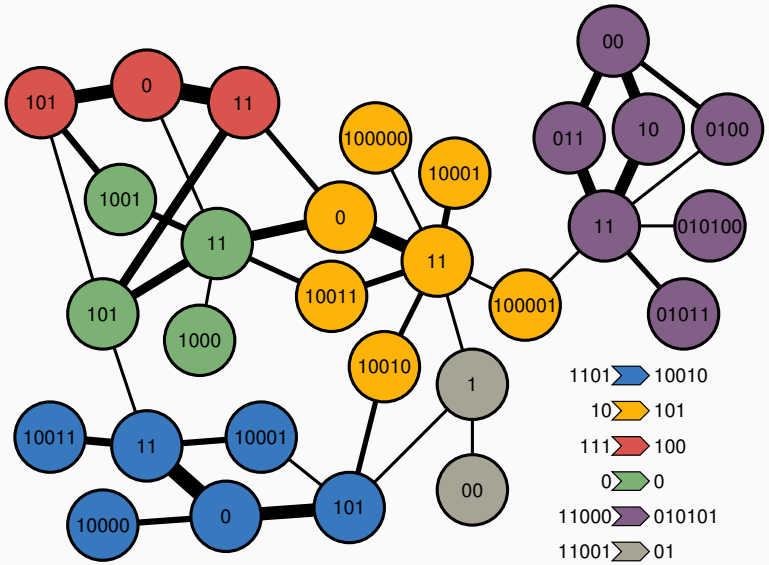
The bipartite map equation

$$L_B(\mathsf{M}) = q^1 H\left(\mathcal{Q}^1\right) + \sum_{i=1}^{m} p_i^1 H\left(\mathcal{P}_i^1\right) + q^2 H\left(\mathcal{Q}^2\right) + \sum_{i=1}^{m} p_i^2 H\left(\mathcal{P}_i^2\right)$$

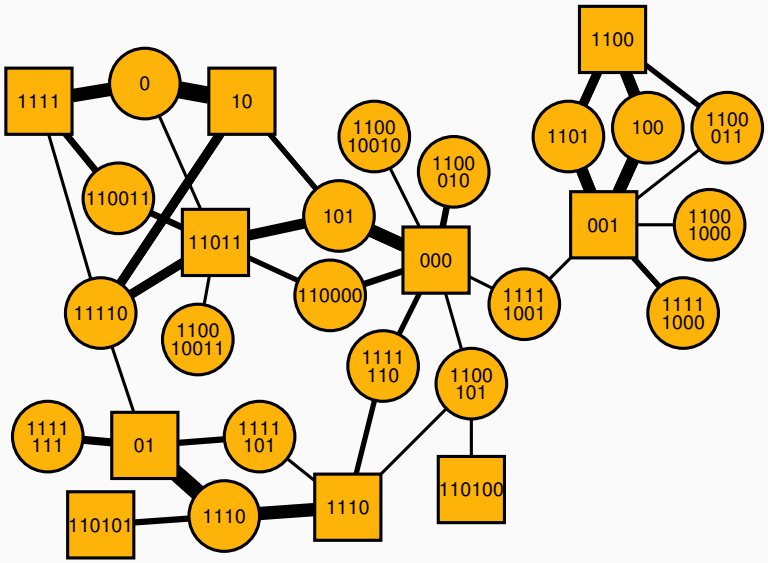
- uses pairs of code books
- alternates between code book types



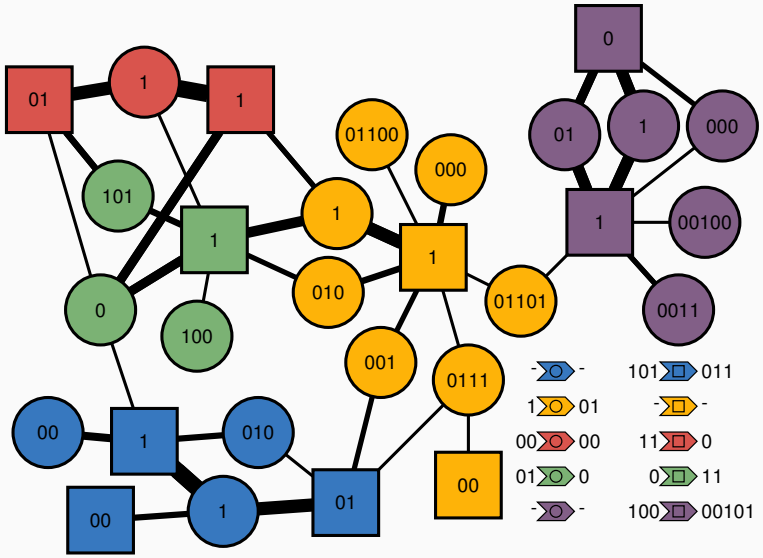
3.85 bits



2.24 bits

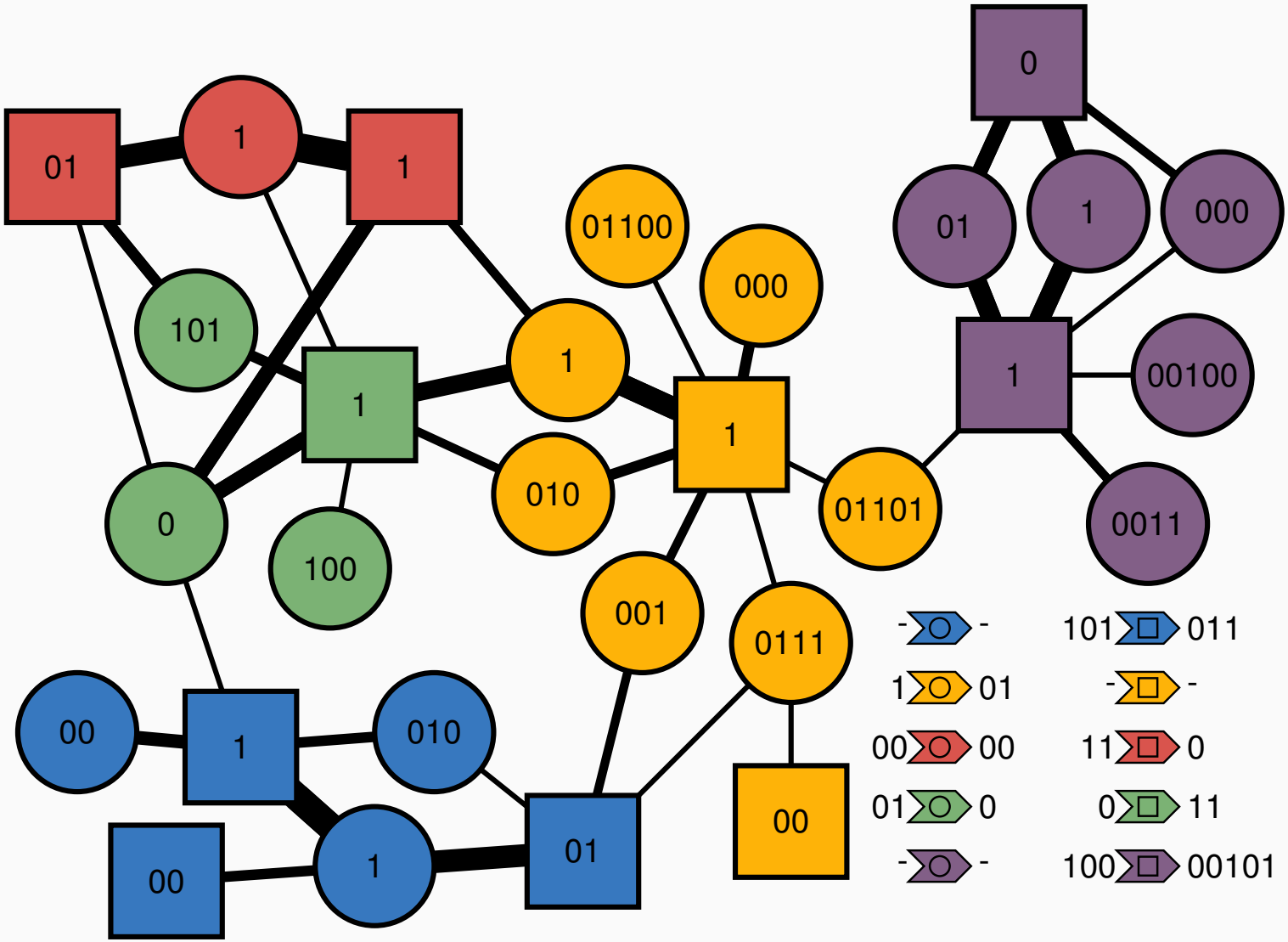


2.85 bits



1.18 bits

What is going on?



Implemented in Infomap (soon public on GitHub)

Name	n_1	n_2	l	L	L_B
IMDb actor-movie	124,414	374,511	1,460,791	11.99	10.88
YouTube user-group	88,490	25,007	286,913	10.25	9.14
GitHub Contest	39,845	99,907	417,361	11.23	10.09
Wiktionary (de)[W]	5,354	144,710	686,661	11.23	10.20
Wiktionary (en)[W]	26,719	2,091,461	5,569,967	12.14	11.08
Virus-host interact.	41	288	433	4.89	3.64
Arroyo Goye[W]	27	8	41	2.70	1.65
Fonseca Ganade[W]	19	10	38	2.24	1.18
Norwegian directors	212	854	1,148	3.83	2.39
Scottish companies	86	131	348	5.17	3.88
Kernel mailing list[W]	34,490	330,155	591,199	9.61	8.54
Last.fm song[W]	1,084,620	992	4,413,834	12.33	11.30
Stack Overflow	524,670	80,492	1,280,982	11.83	10.64
UC Irvine forum[W]	897	520	7,087	8.16	7.14
PGHF meetup	4,989	4,611	39,501	10.90	9.83
SIAM conference	10,018	19	15,533	7.94	6.78
Reuters story-word	19,757	38,677	978,446	13.44	12.40
Wikipedia excellent	2,780	273,959	2,941,902	13.63	12.51

Results

The bipartite map equation

- provides better compression,
- finds more, smaller modules in deeper hierarchies, and
- makes community structures more bipartite.

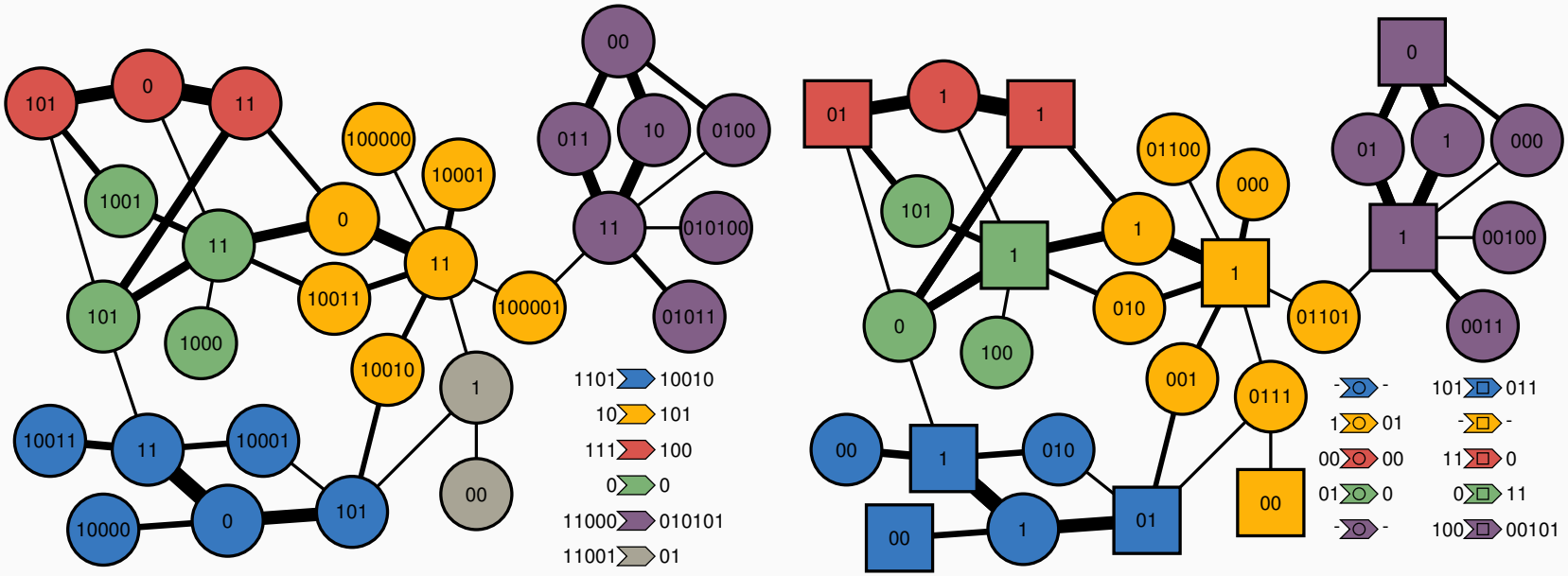
Community Balance

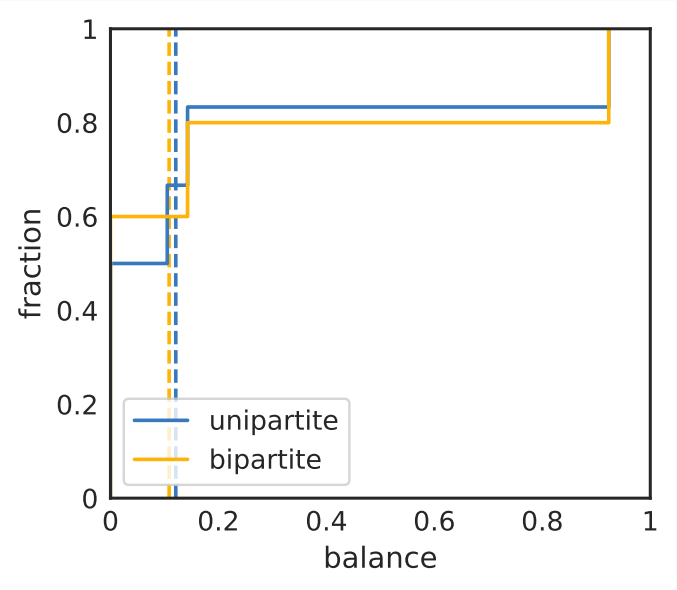
$$b_i = 1 - \left| \frac{q_i^1 - q_i^2}{q_i^1 + q_i^2} \right|$$

Partition Balance

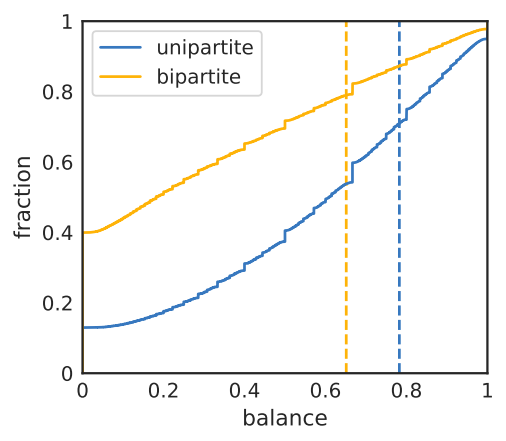
$$b = \sum_i \frac{p_i b_i}{\sum_i p_i}$$

Cumulative distribution of module balance

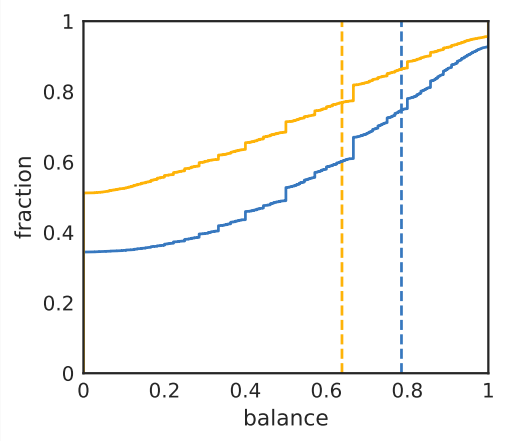




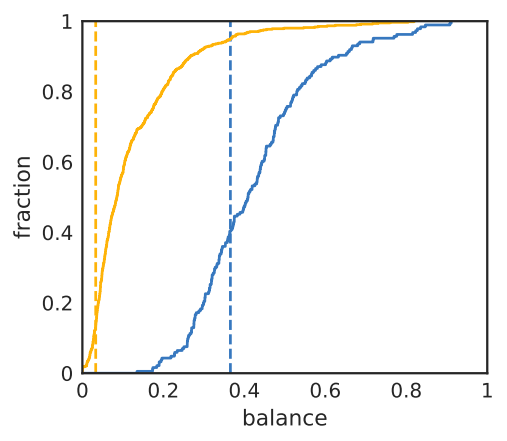
Cumulative distribution of module balance



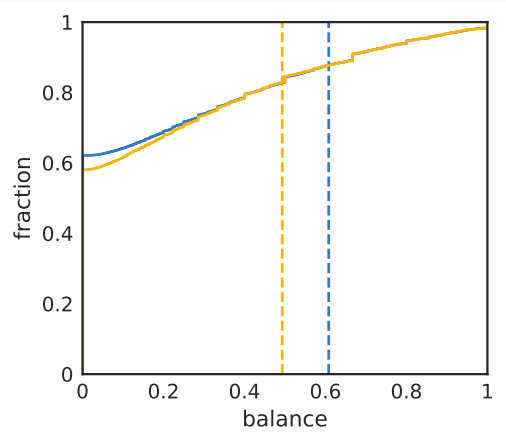
IMDb actor-movie



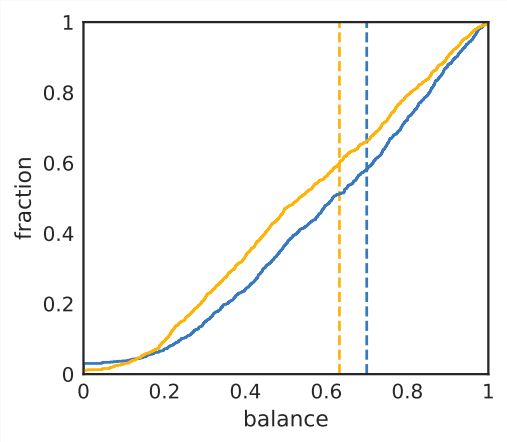
YouTube user-group



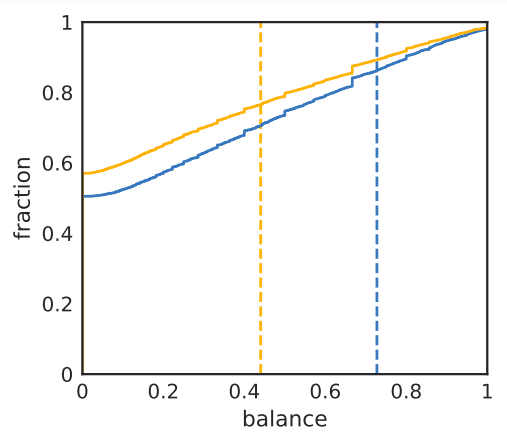
Wikipedia excellent article-word



Wiktionary (en)^W author-article



Last.fm user-song^W



Kernel mailing list^W user-thread

Summary

- Based on the alternating trajectory of random walks in bipartite networks, we use an alternating coding scheme to derive the bipartite map equation
- The bipartite map equation compresses bipartite networks better and finds modules that, presumably, capture the network structure better
- The detected communities are shifted more towards a bipartite connection pattern

arXiv:soon

Thank you for your attention!

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.