Author Contributions Checklist Form

This form documents the artifacts associated with the article (i.e., the data and code supporting the computational findings) and describes how to reproduce the findings.

# Part 1: Data

This paper **does not** involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

## Abstract

Passerine Data: a collection of observations of 63 color-marked passerines, in which individual memberships in groups (individuals appearing together in a flock) were recorded.

Extended Bakery Data: a collection of purchases in a chain of bakery stores, which provide 50 items including 40 bakery goods and 10 drinks. Each purchase contains a collection of items bought together.

## Availability

Data **are** publicly available

Data **cannot be made** publicly available

If the data are publicly available, see the *Publicly available data* section. Otherwise, see the *Non-publicly available dat*a section, below.

### Publicly available data

Data are available online at: https://www.sciencedirect.com/science/article/pii/S0003347215004480

http://users.csc.calpoly.edu/~dekhtyar/466-Spring2018/labs/lab01.html

Data are available as part of the paper’s supplementary material.

Data are publicly available by request, following the process described here:

Data are or will be made available through some other mechanism, described here:

### Non-publicly available data

Discussion of lack of publicly available data:

## Description

### File format(s)

CSV or other plain text:

Software-specific binary format (.Rda, Python pickle, etc.):

Standardized binary format (e.g., netCDF, HDF5, etc.):

Other (described here):

### Data dictionary

Provided by the authors in the following file(s):

Data file(s) is (are) self-describiing (e.g., netCDF files)

Available at the following URL:

<https://www.sciencedirect.com/science/article/pii/S0003347215004480>; http://users.csc.calpoly.edu/~dekhtyar/466-Spring2018/labs/lab01.html

### Additional information (optional)

# Part 2: Code

## Abstract

The R code implements the EM algorithm for estimating the hubs and parameters in the hub model with and without a null component, and the modified EM algorithm for hub set selection in the hub model with an unknown hub set.

## Description

### Code format(s)

Script files

R  Python  Matlab

Other:

Package

R  Python  MATLAB toolbox

Other:

Reproducible report

R Markdown  Jupyter notebook

Other:

Shell script

Other (described here):

### **Supporting software requirements**

Version of primary software used

R version 4.1.0

Libraries and dependencies used by the code

Rsolnp (version 1.16)

### Supporting system/hardware requirements (optional)

### Parallelization used

No parallel code used

Multi-core parallelization on a single machine/node

Number of cores used: 10

Multi-machine/multi-node parallelization

Number of nodes and cores used:

### License

MIT License (default)

BSD

GPL v3.0

Creative Commons

Other (described here):

### Additional information (optional)

# Part 3: Reproducibility workflow

## Scope

The provided workflow reproduces:

Any numbers provided in text in the paper

The computational method(s) presented in the paper (i.e., code is provided that implements the method(s))

All tables and figures in the paper

Selected tables and figures in the paper, as explained and justified here:

## Workflow details

### Format(s)

Single master code file

Wrapper (shell) script(s)

Self-contained R Markdown file, Jupyter notebook, or other literate programming approach

Text file (e.g., a readme-style file) that documents workflow

Makefile

Other (more detail in 'Instructions' below)

### Instructions

Run ‘Onetable.R’, which will produce a txt file, called ‘OneTable.txt’, including table1, table2, table 3 and all data analysis results in the manuscript.

Expected run-time

Approximate time needed to reproduce the analyses on a standard desktop machine:

<1 minute

1-10 minutes

10-60 minutes

1-8 hours

>8 hours

Not feasible to run on a desktop machine, as described here:

### Additional documentation (optional)

# Notes (optional)