Appendix: implementation of the algorithm

Part 1: General description

1. Collection of data
2. Parameterization of contacts between subjects
3. Parameterization of disease natural history
4. Listing and weighting potential transmissions
5. Selection of the likeliest transmissions

Collection of data

To run the WTW algorithm the investigators need to collect:

- a date of onset of symptoms for all observed cases;
- contacts between subjects either at an individual level (e.g. at date *t*, subject *i* had a meeting with individual j during half day) or at an aggregated level (e.g. all staff members of a ward have daily contacts with all patients in the ward).

Parameterization of contacts between subjects

Values must be attributed to each pair of source/infected subjects at each possible date of infection, according to type, length and intensity of contacts between these subjects. When both subjects are in permanent contact during the whole day, the contact value is set to 1; if no contact occurs, the value is set to 0. When information on contact is missing, the value is set to an "uninformative" quantity (e.g. the average value of contacts between this source subject and other subjects at this date).

Contact matrices (one per day of observation) are built with as many lines and columns as subjects in the observed case series. A supplementary line is added for a dummy element representing contacts with subjects not listed in the case series. Its values reflect contacts between unobserved potential sources of infection and subjects in the case series.

Parameterization of disease natural history

Two functions must be defined:

- An “incubation” function, giving the probability per time unit (e.g. day) that a subject was infected at a given date, according to the duration before onset of symptoms.
- An “infectivity” function, giving the probability per unit time that the infectious agent would be transmitted to a susceptible subject according to the time since infection in case of contact of maximal intensity.

Listing and weighting potential transmissions

For each subject in the observed case series, all possible transmissions (i.e. source subjects and dates of infection) are listed. For each possible transmission, a transmission weight is calculated from the product of three values:

- Probability for the infected subject to have been infected at this date, according to his/her date of onset of symptoms (incubation function).
- Infectivity of the source subject at this date, according to her/his date of onset of symptoms (infectivity function). For the dummy element, infectivity can either be a constant value, or vary with time to represent varying intensity of the disease in the community (e.g. for influenza, the infectivity can be based on incidence reported by surveillance systems).
- Contact values between source and infected subjects at this given time.

The transmission weight is defined as minus logarithm of this product (the greater the weight, the less likely the transmission). Note that transmission weights cannot be calculated until the infection date for the source subject is fixed. In this case, the lower bound of the transmission weight is calculated by choosing the maximum possible value of infectivity.

Selection of the likeliest transmissions

We now have, for all cases, a list of possible source subjects, dates of transmission and transmission weights. We build a sequence of transmissions from these observations. A sequence is a subset of all pairwise transmissions. A complete sequence is a sequence considering all subjects. We define the sequence weight as the sum of transmission weights.

1. *Initialization*

We first build a relevant complete sequence of transmission (e.g. by selecting source-infected pairs according to method 2 presented in the paper) and calculate its weight. This complete sequence is set as the “reference sequence.”Any sequence with a higher sequence weight is considered as less likely than the reference sequence and is discarded. Alternatively, the reference sequence weight can be initialized to a large value (note that in this case, computation time can increase dramatically).

2. *Building of sequences with the branch-and-bound method*

The basic principle is that each time a new sequence is built, a sequence weight is calculated and the sequence (and all further sequences including this one) is discarded if its weight is over the “reference sequence” weight.

We first calculate the lower bound of transmission weights for all pairs of source/infected subjects. We select the likeliest transmission, i.e. the sequence with the lowest weight, and incompatible transmissions (e.g. those where source and infected subjects were permuted) are discarded from the set of all remaining possible transmissions

A new pair of source/infected subjects is selected among all remaining transmissions and is added to the sequence. At each step, all incompatible transmissions are discarded. The sequence weight is

calculated and compared to the reference sequence weight. Each time a sequence is discarded, the algorithm moves backwards and modify the last-considered transmission.

3. *Result*

Once all subjects have been considered in a sequence, this sequence is complete and if its weight is below the reference sequence weight, this sequence becomes the new reference sequence.
When all possible transmissions have been evaluated, the algorithm ends and the reference sequence defines the likeliest transmission path.

<u>Part 2</u>: Example

Example: outbreak of influenza in a closed setting with 5 subjects (named A to E), all of whom were infected and symptomatic. We try to retrieve "who infected whom," using to the WTW algorithm.

We assume that the date of onset of symptoms for subjects A to E were on days 6, 8, 10, 12, 8 respectively.

Parameterization of contacts between subjects

We fill the following contact matrices C_{day} (lines: source subjects, columns: infected subjects):

Days 1 to 5:

	A	B	C	D	E
A	NA	0.40	0.30	0.20	0.00
B	0.40	NA	0.35	0.25	0.60
C	0.30	0.35	NA	0.00	0.50
D	0.20	0.25	0.00	NA	0.20
E	0.00	0.60	0.50	0.20	NA
ext	0.10	0.00	0.00	0.20	0.00

Days 6 and 7:

	A	B	C	D	E
A	NA	0.40	0.00	0.20	0.00
B	0.40	NA	0.00	0.25	0.60
C	0.00	0.00	NA	0.00	0.00
D	0.20	0.25	0.00	NA	0.20
E	0.00	0.60	0.00	0.20	NA
ext	0.10	0.00	0.60	0.20	0.00

Days 8 to 12:

	A	B	C	D	E
A	NA	0.40	0.30	0.20	0.00
B	0.40	NA	0.35	0.25	0.30
C	0.30	0.35	NA	0.00	0.25
D	0.20	0.25	0.00	NA	0.10
E	0.00	0.30	0.25	0.10	NA
ext	0.10	0.00	0.00	0.20	0.00

Examples of interpretation for the entries $c_{source,infected,day}$ of these matrices:

- The upper 5x5 submatrices correspond to contacts between subjects in the observed case series, the "ext" lines correspond to the dummy element for contacts with other subjects.
- $c_{A,E,1} = 0$ means that subjects A and E never met each other on day 1.
- $c_{ext,C,1} = 0$ means that subject C never met other subjects than those in the case series on day 1.
- On days 6 and 7, subject C got out of the closed setting
- On days 8 to 12, subject E wore a protection mask, which reduced contacts with any other subject.

Parameterization of natural history for the disease

We define the following incubation function, giving the probability for a subject to have been infected between t+0.5 and t-0.5 days before the onset of her/his symptoms:

```
if t<0 then incub (t) = 0
incub (0) = 0.003
incub (1) = 0.118
incub (2) = 0.551
incub (3) = 0.300
incub (4) = 0.031
incub (5) = 0.005
if t>5 then incub (t) = 0
```

We define the following infectivity function, giving the daily probability that the agent would be transmitted from an infectious to a susceptible subject according to the time since infection.

```
if t<0 then infec (t) = 0
infec (0) = 0.005
infec (1) = 0.110
infec (2) = 0.240
infec (3) = 0.247
infec (4) = 0.179
infec (5) = 0.107
infec (6) = 0.058
infec (7) = 0.025
infec (8) = 0.011
infec (9) = 0.007
infec (10) = 0.002
infec (11) = 0.001
if t>12 then infec (t) = 0
```

Listing and weighting of potential transmissions

For all dates of follow-up and for each couple of source and infected subjects, we define a transmission weight:

$$t_{source,infected,day} = -\log[\, infec\ (day - date_{infection\_source}) \times c_{source,infected,day} \times incub\ (date_{onset\_of\_symptoms\_infected} - day)\ ]$$

To be calculated, this transmission weight needs making hypotheses on date of infection of source subjects. The lower bound of this weight can be calculated until such hypotheses have been made:

$$lower\_bound\_t_{source,infected,day} = -\log[\ 0.247 \times c_{source,infected,day} \times incub\ (date_{onset\_of\_symptoms\_infected} - day)\ ]$$

To launch the branch and band algorithm, we calculate the lower limit of all transmission weights.

Selection of the likeliest transmissions

Initialization

We set the initial reference sequence weight to $+\infty$.

1st transmission

According to the lower bounds of transmission weights, the overall numbers of possible transmissions for subjects A, B, C, D and E are 23, 22, 14, 24 and 16 respectively.

We first study possible source subjects and dates for infection of subject C. C may have been infected by:

- A on days 5, 8, 9 or 10
- B on days 5, 8, 9 or 10
- D on days 5, 8, 9 or 10
- E on days 5, 8, 9 or 10
- ext on days 6 and 7

We choose the transmission "B infected C on day 8" (noted BC8), for its lower bound is the lowest. Other possible transmissions will be studied later.

We can now discard all irrelevant sequences according to this assumption, i.e. sequences in which:

- any other source subject than B infected C;
- or C was infected any other day than day 8;
- or a source subject infected B after day 8;
- or C infected another subject before day 8.

BC8 is the beginning of a sequence whose weight is still unknown but over 3.044 (lower bound of the weight of transmission BC8).

Transmission	Weight	Lower bound of weight
BC8	Unknown	3.044

Lower bound of weight for sequence BC8: 3.044

2nd transmission

The remaining numbers of possible transmissions for subjects A, B, D and E are 18, 18, 24 and 13. E is now chosen.

The retained transmission to E is BE6. Irrelevant sequences are discarded.

Transmission	Weight	Lower bound of weight
BC8	Unknown	3.044
BE6	Unknown	2.505

Lower bound of weight for sequence BC8_BE6: 3.044 + 2.505 = 5.549

3rd transmission

The remaining numbers of possible transmissions for subjects A, B and D are 18, 8 and 24. B is now chosen.

The retained transmission to B is AB6. Irrelevant sequences are discarded. As previously, the weight of transmission AB6 cannot be calculated yet. However, setting a date of infection for subject B allows to calculate the infectivity for this subject, and hereafter the weights of BC8 and BE6:

Transmission	Weight	Lower bound of weight
BC8	3.074	3.074
BE6	6.410	6.410
AB6	unknown	2.910

Lower bound of weight for sequence BC8_BE6_AB6 = 3.074 + 6.410 + 2.910 = 12.394

4th transmission

The remaining numbers of possible transmissions for subjects A and D are 12 and 24. A is now chosen.

The retained transmission to A is DA4. After discarding irrelevant sequences, we note that no solution remains for infection of D. This means that the sequence BC8_BE6_AB6_DA4 is irrelevant and we must move backwards to change the last-considered transmission (DA4).

Regarding other possible transmissions to A, the weight whose lower bound is the lowest is DA6. We retain this transmission. After discarding irrelevant sequences, we note once again that no solution remains for infection of D. We move backwards again to change the last-considered transmission (DA6).

Regarding other possible transmissions to A, the weight whose lower bound is the lowest is extA4 (infection from the dummy element for subjects not listed in the case series, on day 4). We retain this transmission and discard irrelevant sequences.

Transmission	Weight	Lower bound of weight
BC8	3.074	3.074
BE6	6.410	6.410
AB6	2.940	2.940
extA4	5.201	5.201

Weight of sequence BC8_BE6_AB6_extA4 = 3.074 + 6.410 + 2.940 + 5.201 = 17.625

5th transmission

The last subject whose infection needs being retrieved is D. Among the 24 possible transmissions, the one whose weight is the lowest (BD10) is retained. Note that since transmissions have been proposed for infections of all subjects, this sequence is complete and its real weight can be calculated.

Weight of sequence BC8_BE6_AB6_extA4_BD10 = 3.074 + 6.410 + 2.940 + 5.201 + 3.7021= 21.370

This first complete sequence becomes the reference sequence. From now on, any sequence whose weight is over 21.370 will be safely discarded.

Continuation of the algorithm

In our example:

- BC8_BE6_AB6_extA4_BD10 is the best sequence relying on BC8_BE6_AB6_extA4. All other possible transmissions to D are discarded and we move backwards to propose a new transmission to the previous subject (A).
- The next likeliest transmission to A is extA3 (weight of BC8_BE6_AB6_extA3: 18.233).

- Among the 24 remaining possible transmissions to D, the minimal weight is 3.702. This means that the weight of any sequence relying on BC8_BE6_AB6_extA3 will be at least 18.233+3.702=21.935, i.e. over the reference sequence weight.
- After evaluating all possible sequences including BC8_BE6_AB6, we note that all sequences will also lead to suboptimal results. We therefore change the last-considered transmission and move backwards to evaluate BC8_BE6_DB6, and so on.

The branch-and-bound algorithm ends when all possible sequences have been evaluated. The likeliest transmission path is the one described by the latest reference sequence.