Some Scaling Properties of Traffic in Communication Networks

Paulo Gonçalves
DANTE - INRIA - ENS Lyon

Patrick Loiseau (PhD, 2006-2009)
Shubhabrata Roy (PhD, 2010-2013)
M. Sokol (PhD, 2010- )
B. Girault (PhD, 2012-2015)

Seminars Complex Networks, LIP6, UPMC – march 7, 2013
Historical perspective

- 1917: Erlang - Circuit switching networks
- 1969: Kleinrock - Packet switching networks
- 1988: Van Jacobson - AIMD - TCP
- 1992: Tim Berners Lee - Web
- 1993: Leland Willinger - Taqqu - LRD LAN
- 1994: Paxson Floyd - LRD WAN
- 1992: Crovella - Heavy tails
- 1997: Taqqu - ON/OFF model
- 1993: Leland Willinger - Taqqu - LRD LAN

Some open questions:
- Long Range Dependence / Heavy Tailed distributions impact on QoS?
- Existing models (e.g. Padhye) only predict mean metrics (e.g. throughput): what about variability?
Historical perspective

1917 Erlang: circuit switching networks
1969 Kleinrock: packet switching networks
1988 Van Jacobson: AIMD - TCP
1992 Tim Berners Lee: Web
1993 Leland Willinger
1994 Paxson Floyd: LRD WAN
1994 Norros: queues and LRD
1997 Taqqu: LRD LAN
1997 Crovella: heavy tails
1997 Park Crovella: QoS degradation
1998 Padhye: Markov - 1 TCP source
2004 Roberts: QoS insensitivity
2004 Mandjes: QoS ON/OFF (theo.)

exponential
Poisson
Markov
1968-69
Mandelbrot
LRD (fBm)
heavy tails
Historical perspective

- 1917 Erlang: circuit switching networks
- 1969 Kleinrock: packet switching networks
- 1988 Van Jacobson: AIMD - TCP
- 1992 Tim Berners Lee: Web
- 1994 Paxson Floyd: LRD WAN
- 1994 Norros: queues and LRD
- 1997 Taqqu: heavy tails
- 1997 Park Crovella: QoS degradation
- 1998 Padhye: Markov - 1 TCP source
- 2004 Roberts: QoS insensitivity
- 2004 Mandjes: QoS ON/OFF (theo.)

Some open questions:
- Long Range Dependence / Heavy Tailed distributions impact on QoS?
- Existing models (e.g. Padhye) only predict mean metrics (e.g. throughput): what about variability?
Our approach

To combine theoretical models with controlled experiments in realistic environments and real-world traffic traces
Simplified System

- Congestion essentially arises at the access points
  → Simplified System: single bottleneck
- Users’ behavior: ON/OFF source model
- *MetroFlux*: a probe for traffic capture at packet level (O. Goga, …)
Long memory in aggregated traffic: the Taqqu model

- Heavy-tailed distributed ON periods: heavy tail index $\alpha_{ON} > 1$

Theorem (Taqqu, Willinger, Sherman, 1997)

In the limit of a large number of sources $N_{src}$, if:

- flow throughput is constant,
- same throughput for all flows;

aggregated bandwidth $B^{(\Delta)}(t)$ is long range dependent, with parameter:

$$H = \max \left( \frac{3 - \alpha_{ON}}{2}, \frac{1}{2} \right)$$

Long memory: long range correlation ($H > 1/2$)

$$\text{Cov}_{B^{(\Delta)}}(\tau) = \mathbb{E} \left\{ B^{(\Delta)}(t)B^{(\Delta)}(t + \tau) \right\} \sim \tau^{(2H-2)}$$

Variance grows faster than $\Delta$: $\text{Var} \left\{ B^{(\Delta)}(t) \right\} \sim \Delta^{2H}$
Theorem validation on a realistic environment

- Controlled experiment: *MetroFlux* 1 Gbps, 100 sources, 8 hours traffic
- UDP/TCP: throughput limited to 5 Mbps (no congestion)

- ON Distribution (source)
- Log-diagram (aggregated traffic)
- Taqqu Prediction

⇒ Protocol has no influence at large scales
⇒ Long memory shows up beyond scale $\Delta = \mu_{ON}$ (mean flow duration)
Influence of flow mean throughput / duration correlation

- Web traffic acquired at in2p3 (Lyon) with *MetroFlux* 10 Gbps

**ON Distribution**

<table>
<thead>
<tr>
<th>Duration (ms)</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>10</td>
</tr>
<tr>
<td>0.1</td>
<td>10^{-5}</td>
</tr>
<tr>
<td>1</td>
<td>10^{-10}</td>
</tr>
<tr>
<td>10</td>
<td>10^{-15}</td>
</tr>
<tr>
<td>100</td>
<td>10^{-20}</td>
</tr>
<tr>
<td>1000</td>
<td>10^{-25}</td>
</tr>
</tbody>
</table>

\[ \alpha_{ON} = 1.2 \]

**Size Distribution**

<table>
<thead>
<tr>
<th>Size (Gbps)</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>10^{-10}</td>
</tr>
<tr>
<td>0.1</td>
<td>10^{-5}</td>
</tr>
<tr>
<td>1</td>
<td>10^{-1}</td>
</tr>
<tr>
<td>10</td>
<td>10^{1}</td>
</tr>
<tr>
<td>100</td>
<td>10^{5}</td>
</tr>
<tr>
<td>1000</td>
<td>10^{10}</td>
</tr>
</tbody>
</table>

\[ \alpha_{SI} = 0.85 \]

- Heavy-tailed ON periods, \( \alpha_{ON} = 1.2 \)
- Heavy tailed flow sizes, \( \alpha_{SI} = 0.85 \)
Influence of flow mean throughput / duration correlation

- Web traffic acquired at in2p3 (Lyon) with MetroFlux 10 Gbps

- Heavy-tailed ON periods, $\alpha_{ON} = 1.2$

- Heavy-tailed flow sizes, $\alpha_{SI} = 0.85$

- Flow throughput and duration are correlated:

  $$\mathbb{E}\{\text{thr.} | \text{dur.}\} \propto (\text{dur.})^{\beta-1}, \quad \beta = \frac{\alpha_{ON}}{\alpha_{SI}} \quad (= 1.4)$$

  ⇒ Which heavy tail index does control LRD? ($\alpha_{ON}$, $\alpha_{SI}$)?
Taqqu model extension

- Planar Poisson process to describe arrival instant vs duration

![Diagram showing the Taqqu model extension](image)
Taqqu model extension

- Planar Poisson process to describe arrival instant vs duration

**Proposition (LGVBP, 2009)**

**Model:** $\mathbb{E}\{\text{through} \mid \text{dur.}\} = M \cdot (\text{dur.})^{\beta - 1}$; $\mathbb{V} \mathbb{A} \mathbb{R}\{\text{through} \mid \text{dur.}\} = V$

$$\mathbb{C} \mathbb{O} \mathbb{v}_{B(\Delta)}(\tau) = CM^2\tau^{-(\alpha_{ON} - 2(\beta - 1)) + 1} + C'V\tau^{-\alpha_{ON} + 1}$$
Taqqu model extension

- Planar Poisson process to describe arrival instant vs duration

**Proposition (LGVBP, 2009)**

**Model:** $\mathbb{E}\{\text{through.}|\text{dur.}\} = M \cdot (\text{dur.})^{\beta-1}$; $\mathbb{V}\text{ar}\{\text{through.}|\text{dur.}\} = V$

$$\text{Cov}_{B(\Delta)}(\tau) = CM^2 \tau^{-(\alpha ON - 2(\beta-1)) + 1} + C' V \tau^{-\alpha ON + 1}$$

Log-diagram, $\beta > 1$

- threshold $\tau^* = \left(\frac{C' V}{CM^2}\right)^{1/(2(\beta-1))}$
Scaling Properties of Traffic

Heavy tailed distributions and long range dependence

Taqqu model extension

• Planar Poisson process to describe arrival instant vs duration

Proposition (LGVBP, 2009)

Model:

\[ E\{ \text{through.|dur.} \} = M \cdot (\text{dur.})^{\beta - 1}; \quad \text{Var}\{ \text{through.|dur.} \} = V \]

\[ \text{Cov}_{B(\Delta)}(\tau) = CM^2 \tau^{-(\alpha_{ON} - 2(\beta - 1)) + 1} + C' V \tau^{-\alpha_{ON} + 1} \]

Log-diagram, \( \beta > 1 \)

\[ \text{H=H}_\text{Taqqu} + (\beta - 1) \]

threshold \( \tau^* = \left( \frac{C' V}{CM^2} \right)^{1/(2(\beta - 1))} \)

\[ \rightarrow \text{if} \ \Delta \gg \tau^*: \ H = H_\text{Taqqu} + (\beta - 1) \]
Taqqu model extension

- Planar Poisson process to describe arrival instant vs duration

Proposition (LGVBP, 2009)

Model: \( \mathbb{E}\{\text{through.}|\text{dur.}\} = M \cdot (\text{dur.})^{\beta - 1} \); \( \mathbb{V} \{\text{through.}|\text{dur.}\} = V \)

\[
\text{Cov}_{\mathcal{B}(\Delta)}(\tau) = CM^2 \tau^{-(\alpha ON - 2(\beta - 1)) + 1} + C'V^\tau^{-\alpha ON + 1}
\]

Log-diagram, \( \beta > 1 \)

- threshold \( \tau^* = \left( \frac{C'V}{CM^2} \right)^{1/(2(\beta - 1))} \)

\[ \text{if } \Delta \gg \tau^* : H = H_{\text{Taqqu}} + (\beta - 1) \]

\[ \text{if } \Delta \ll \tau^* : H = H_{\text{Taqqu}} \]
Taqqu model extension

- Planar Poisson process to describe arrival instant vs duration

**Proposition (LGVBP, 2009)**

Model: $E\{\text{through.}|\text{dur.}\} = M \cdot (\text{dur.})^{\beta-1}$; $Var\{\text{through.}|\text{dur.}\} = V$

$$\text{Cov}_{B(\Delta)}(\tau) = CM^2 \tau^{-(\alpha ON - 2(\beta-1)) + 1} + C' V \tau^{-\alpha ON + 1}$$

Log-diagram, $\beta > 1$

- Correlations intensify LRD ($\beta > 1$)
- Traffic evolution, future Internet: “flow-aware” control mechanisms, FTTH
LRD impact on QoS: a brief (experimental) outlook

The situation is complex...
LRD impact on QoS: a brief (experimental) outlook

The situation is complex...

- Negative on finite queues with UDP flows [cf. Mandjes, 2004 (infinite queues)]
  - LRD degrades QoS for large queue sizes (beyond some threshold)
  - but the threshold depends on the considered QoS metric (loss rate vs mean load)

  - LRD has contradictory effects on QoS metrics depending on:
    - slow start without slow start
      - Delay $\downarrow \uparrow$
      - loss rate $\downarrow \rightarrow$
      - mean throughput $\rightarrow \uparrow$
LRD impact on QoS: a brief (experimental) outlook

The situation is complex...

- Negative on finite queues with UDP flows [cf. Mandjes, 2004 (infinite queues)]
  - LRD degrades QoS for large queue sizes (beyond some threshold)
  - but the threshold depends on the considered QoS metric (loss rate vs mean load)

  - LRD has contradictory effects on QoS metrics depending on:

    | with slow start | without slow start |
    |-----------------|--------------------|
    | Delay           | ↘                  |
    | loss rate       | ↘                  |
    | mean throughput | →                  |

- Heavy tailed distributions (i.e LRD) can favour QoS for large flows
LRD impact on QoS: a brief (experimental) outlook

The situation is complex...

- Negative on finite queues with UDP flows [cf. Mandjes, 2004 (infinite queues)]
  - LRD degrades QoS for large queue sizes (beyond some threshold)
  - but the threshold depends on the considered QoS metric (loss rate vs mean load)

  - LRD has contradictory effects on QoS metrics depending on:

<table>
<thead>
<tr>
<th></th>
<th>with slow start</th>
<th>without slow start</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay</td>
<td>↘</td>
<td>↗</td>
</tr>
<tr>
<td>loss rate</td>
<td>↘</td>
<td>→</td>
</tr>
<tr>
<td>mean throughput</td>
<td>→</td>
<td>↗</td>
</tr>
</tbody>
</table>

- Heavy tailed distributions (i.e LRD) can favour QoS for large flows

- But in general, QoS is a complex function of multiple variables
Second level of description: single TCP source traffic

Sources

\[ \begin{array}{c}
1 \\
\cdot \\
\cdot \\
\cdot \\
N_{src}
\end{array} \]

Agrégat

\[ \begin{array}{c}
\text{\,} \\
\text{T}_{ON} \\
\text{T}_{OFF}
\end{array} \]
Second level of description: single TCP source traffic

- Single TCP source traffic detail
- Long-lived flow $\rightarrow$ stationary regime

$\Rightarrow$ How to characterize the congestion window evolution?
Markov model

$W_i$ (paquets)

- long-lived flow stationary regime: AIMD
- model: \((W_i)_{i \geq 1}\) finite Markov chain (irreducible, aperiodic), transition matrix \(Q\):

\[
\begin{align*}
Q_{w,\min(w+1,w_{\max})} &= 1 - p(w), \\
Q_{w,\max(\lfloor w/2 \rfloor,1)} &= p(w).
\end{align*}
\]

- \(p(\cdot)\) loss probability of at least one packet, only depends on the current congestion window (hyp.)
- Example: [Padhye, 1998] Bernoulli loss: \(p(w) = 1 - (1 - p_{pkt})^w\)
Almost sure mean throughput

\[ \overline{W}_i^{(n)} (\text{paquets}) \]

\[ \overline{W}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} W_i \]

- Mean throughput at scale \( n \) (RTT):

Ergodic Birkhoff theorem (1931): almost sure mean

For almost all realisation, the mean throughput at scale \( n \) converges towards a value corresponding to the expectation of the invariant distribution:

\[ \overline{W}^{(n)} \xrightarrow{p.s.} \overline{W}^{(\infty)} = E\{W_i\} \]

- Example: [Padhye, 1998], \( \overline{W}^{(\infty)} \sim \frac{3}{2p_{\text{pkt}}} \) \( p_{\text{pkt}} \to 0 \) (RTT=1, MSS=1)
Throughput variability: Large Deviations

- \( \overline{W}^{(n)} \approx \alpha \neq \overline{W}^{(\infty)} \) Rare events

**Large Deviations theorem (Ellis, 84)**

\[
P(\overline{W}^{(n)} \approx \alpha) \sim \exp(n \cdot f(\alpha))
\]

- \( f(\alpha) \) Large Deviation spectrum
- \( \rightarrow \) Scale invariant quantity
Throughput variability: Large Deviations

- $\overline{W}^{(n)} \simeq \alpha \neq \overline{W}^{(\infty)}$ Rare events

Large Deviations theorem (Ellis, 84)

$$\mathbb{P}(\overline{W}^{(n)} \simeq \alpha) \underset{n \to \infty}{\sim} \exp(n \cdot f(\alpha))$$

- $f(\alpha)$ Large Deviation spectrum

→ Scale invariant quantity

⇒ Does a similar theorem exist for a single realization?
Large Deviation on almost all realizations

For a given $\alpha$, if $k_n \geq e^{nR(\alpha)}$, then a.s.

$$\# \left\{ j \in \{1, \cdots, k_n\} : \overline{W}_j^{(n)} \simeq \alpha \right\} \sim \frac{\exp(n \cdot f(\alpha))}{k_n}$$

- “Price to pay”: exponential increase of the number of intervals
- Finite realization (of size $N$): $nk_n = N$

$[\alpha_{\text{min}}(n), \alpha_{\text{max}}(n)]$ support of observable spectrum at scale $n$
Large Deviation on almost all realizations

For a given $\alpha$, if $k_n \geq e^{nR(\alpha)}$, then a.s.

$$\frac{\# \left\{ j \in \{1, \cdots, k_n\} : \overline{W}_j^{(n)} \simeq \alpha \right\}}{k_n} \sim \exp(n \cdot f(\alpha))$$

“Price to pay”: exponential increase of the number of intervals

Finite realization (of size $N$): $nk_n = N$

$[\alpha_{\min}(n), \alpha_{\max}(n)]$ support of observable spectrum at scale $n$

- Theory: $p(\cdot) \rightarrow Q \rightarrow f(\alpha), R(\alpha), \alpha_{\min}, \alpha_{\max}$
- Practice: $(W_i)_{i \leq N} \rightarrow$ observed distribution
Results: example of Bernoulli losses ($p_{\text{pkt}} = 0.02$)
Results: example of Bernoulli losses ($p_{pkt} = 0.02$)

- Apex: almost sure mean: 8.6 packets (Padhye: $\sqrt{\frac{3}{2p_{pkt}}} = 8.66$)
Results: example of Bernoulli losses ($p_{pkt} = 0.02$)

- Apex: almost sure mean: 8.6 packets (Padhye: $\sqrt{\frac{3}{2p_{pkt}}} = 8.66$)
Results: example of Bernoulli losses ($p_{\text{pkt}} = 0.02$)

Apex: almost sure mean: 8.6 packets (Padhye: $\sqrt{\frac{3}{2p_{\text{pkt}}}} = 8.66$)

$W(\infty)$
Results: example of Bernoulli losses \( (p_{\text{pkt}} = 0.02) \)

- Apex: almost sure mean: 8.6 packets (Padhye: \( \sqrt{\frac{3}{2p_{\text{pkt}}}} = 8.66 \))
- Superimposition at different scales \( \rightarrow \) scale invariance
Results: example of Bernoulli losses \((p_{pt} = 0.02)\)

- Apex: almost sure mean: 8.6 packets (Padhye: \(\sqrt{\frac{3}{2p_{pkt}}} = 8.66\))
- Superimposition at different scales \(\rightarrow\) scale invariance
- Beyond \(n = 100\): variability
  - \(n = 100\), portion of intervals with mean \(\sim 11\): \(e^{-100 \times 0.01} = 0.37\)
  - \(n = 200\), portion of intervals with mean \(\sim 11\): \(e^{-200 \times 0.01} = 0.14\)

\(\Rightarrow\) More accurate information than the almost sure mean
Results II: case of a long-lived flow

- losses: not Bernoulli
- empirical losses
Two important assets for Large Deviations Utility

General result ("Large deviations for the local fluctuations of random walks", J. Barral, P. Loiseau, Stochastic Processes and their Applications, 2011)

A wide class of processes (stationary & mixing) verifies an empirical large deviation principle. In particular, this results holds true any time series that can reliably be modelled by an irreducible, aperiodic Markov process.


We derived a consistent estimator of the large deviation spectrum from a finite size time series (observation samples). We proved convergence on mathematical objects with scale invariance properties (multifractal measures and processes).
An epidemic based model for volatile workload

Goal – Dynamic resource allocation yielding a good compromise between capex and opex costs
An epidemic based model for volatile workload

**Goal** – Dynamic resource allocation yielding a good compromise between capex and opex costs

**Approach** – Combine the three ingredients:
- A sensible (epidemic) model to catch the burstiness and the dynamics of the workload
- A (Markov) model that verifies a large deviation principle
- A probabilistic management policy based on the large deviation characterisation
An epidemic based model for volatile workload

**Goal** – Dynamic resource allocation yielding a good compromise between capex and opex costs

**Approach** – Combine the three ingredients:

- A sensible (epidemic) model to catch the burstiness and the dynamics of the workload
- A (Markov) model that verifies a large deviation principle
- A probabilistic management policy based on the large deviation characterisation

![Number of current VoD users over time]
An epidemic based model for volatile workload

**Goal** – Dynamic resource allocation yielding a good compromise between capex and opex costs

**Approach** – Combine the three ingredients:
- A sensible (epidemic) model to catch the burstiness and the dynamics of the workload
- A (Markov) model that verifies a large deviation principle
- A probabilistic management policy based on the large deviation characterisation

A hidden state Markov process with memory effect

\[ \begin{align*}
\beta(i+r)+l & \\
\mu r & \\
\gamma i & \\
\end{align*} \]

\[ \begin{align*}
i, r & \\
i, r-1 & \\
i+1, r & \\
i-1, r+1 & \\
\end{align*} \]

\[ \begin{align*}
\beta & = \beta_1 \\
\beta & = \beta_2 \\
\end{align*} \]

\[ i : \text{current } \# \text{ of viewers} \quad r : \text{current } \# \text{ of infected} \]
An epidemic based model for volatile workload

Calibration and evaluation

VoD workload trace

Memory Markov model

Modul. Markov Poisson

Steady state distribution

Autocorrelation function

Param. estimation precision
Markov processes

Under mild conditions, a Markov processes $I_t$ verifies a large deviation principle:

$$
P\{\langle I_t \rangle^\tau \approx \alpha \} \equiv \exp (\tau \cdot f(\alpha)), \quad \tau \to \infty$$

$f(\alpha)$ : theoretically (from the transition matrix) or empirically (from a finite trace) identifiable
Markov processes

Under mild conditions, a Markov processes $l_t$ verifies a large deviation principle:

$$\mathbb{P}\{\langle l_t \rangle_\tau \approx \alpha \} \equiv \exp(\tau \cdot f(\alpha)), \quad \tau \to \infty$$

$f(\alpha)$: theoretically (from the transition matrix) or empirically (from a finite trace) identifiable
Markov processes

Under mild conditions, a Markov processes $I_t$ verifies a large deviation principle:

$$\mathbb{P}\{\langle I_t \rangle_\tau \approx \alpha \} \equiv \exp (\tau \cdot f(\alpha)), \quad \tau \to \infty$$

$f(\alpha)$: theoretically (from the transition matrix) or empirically (from a finite trace) identifiable

"Dynamic" implies time scale: a notion that is explicit in large deviation principle
Parametric generalisation of semi-supervised learning

**Standard classification**

*Training set*

\[(X^{(t)}, Y^{(t)}) \rightarrow \text{classifier}\]

*Validation set*

\[X^{(v)} \xrightarrow{\text{classifier}} Y^{(v')} : |Y^{(v)} - Y^{(v')}| \approx 0\]

*Real data*

\[X \xrightarrow{\text{classifier}} \text{Answer}\]

**Semi-supervised classification**

*Validation set*

\[(X^{(v)}, Y^{(L)}) \xrightarrow{\text{classifier}} Y^{(v')}\]

such that \[|Y^{(v)} - Y^{(v')}| \approx 0\]

*Real data*

\[(X, Y^{(L)}) \xrightarrow{\text{classifier}} \text{Answer}\]

Allow to constantly update the classifier to match data evolution
Parametric generalisation of semi-supervised learning

**Standard classification**

**Training set**

\[(X^{(t)}, Y^{(t)}) \rightarrow \text{classifier}\]

**Validation set**

\[X^{(v)} \rightarrow Y^{(v')} : |Y^{(v)} - Y^{(v')}| \simeq 0\]

**Real data**

\[X \rightarrow \text{Answer}\]

**Semi-supervised classification**

**Validation set**

\[(X^{(v)}, Y^{(L)})^{\text{classifier}} Y^{(v')}
\]

\[\text{such that } |Y^{(v)} - Y^{(v')}| \simeq 0\]

**Real data**

\[(X, Y^{(L)})^{\text{classifier}} \text{Answer}\]

Allow to constantly update the classifier to match data evolution

**Dataset**

\[X = X_1, X_2, \ldots, X_p, X_{p+1}, \ldots, X_N\]

labeled points

**Similarity matrix**

\[W\]

and \[D\] (reap. \[D^*\]) the row-sum (reap. column) of \[W\]

**Label matrix**

\[Y = \{Y_{i,k} \in (0,1) \text{ for } i = 1, \ldots N \text{ and } k = 1, \ldots K\}\]

**Objective (classification) matrix**

\[F_{N \times K} : \text{element } i \text{ belongs to class } k^* = \arg\max_k F_{i,k}\]
Parametric generalisation of semi-supervised learning

Standard Laplacian solution

\[
\arg\max_F \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \| F_i. - F_j. \|^2 + \mu \sum_{i=1}^{N} d_i \| F_i. - Y_i. \| ^2 \right\}
\]
Parametric generalisation of semi-supervised learning

Standard Laplacian solution

$$\arg\max_{F} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \| F_i. - F_j. \|^2 + \mu \sum_{i=1}^{N} d_i \| F_i. - Y_i. \|^2 \right\}$$

Generalised semi-supervised classification [M. Sokol, 2012]

$$\arg\max_{F} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \| d_i^{\sigma-1} F_i. - d_j^{\sigma-1} F_j. \|^2 + \mu \sum_{i=1}^{N} d_i^{2\sigma-1} \| F_i. - Y_i. \|^2 \right\}$$
Parametric generalisation of semi-supervised learning

Standard Laplacian solution

$$\arg\max_F \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \| F_i - F_j \|^2 + \mu \sum_{i=1}^{N} d_i \| F_i - Y_i \|^2 \right\}$$

Generalised semi-supervised classification [M. Sokol, 2012]

$$\arg\max_F \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \| d_i^{\sigma-1} F_i - d_j^{\sigma-1} F_j \|^2 + \mu \sum_{i=1}^{N} d_i^{2\sigma-1} \| F_i - Y_i \|^2 \right\}$$

- $\sigma = 1$  Standard Laplacian  (Random walk from unlabelled to labelled points)
- $\sigma = 1/2$  Normalised Laplacian
- $\sigma = 0$  PageRank method  (Random walk from labelled to unlabelled points)
Parametric generalisation of semi-supervised learning

Standard Laplacian solution

$$\arg\max_{F} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \| F_i - F_j \|^{2} + \mu \sum_{i=1}^{N} d_i \| F_i - Y_i \|^{2} \right\}$$

Generalised semi-supervised classification [M. Sokol, 2012]

$$\arg\max_{F} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \| d_{i}^{\sigma-1} F_i - d_{j}^{\sigma-1} F_j \|^{2} + \mu \sum_{i=1}^{N} d_{i}^{2\sigma-1} \| F_i - Y_i \|^{2} \right\}$$

$\sigma = 1$  
Standard Laplacian  
(Random walk from unlabelled to labelled points)

$\sigma = 1/2$  
Normalised Laplacian  
(Random walk from labelled to unlabelled points)

$\sigma = 0$  
PageRank method

$$F_{.k} = \frac{\mu}{2 + \mu} \left( I - \frac{2}{2 + \mu} D^{-\sigma} WD^{\sigma-1} \right)^{-1} Y_{.k}, \text{ for } k = 1, \ldots, K$$

Tune the value of parameter $\sigma$ to match the dataset
Duality and semi-supervised learning

graph (similarity) \(\leftrightarrow\) process (metric)

formulation (multidimensional scaling): \[
\arg\max_F \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} (\|F_i - F_j\| - w_{ij})^2 \right\}
\]
Duality and semi-supervised learning

\[ \text{argmax}_{F} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \| F_i - F_j \| - w_{ij} \right)^2 \right\} \]

- bridge ordination (MDS) and generalised semi-supervised learning
  - leverage \( \sigma \) flexibility to vary duality principle
Duality and semi-supervised learning

graph (similarity) \leftrightarrow \text{process (metric)}

formulation (\textit{multidimensional scaling}): \arg\max_{F} \left\lbrace \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \| F_i - F_j \| - w_{ij} \right)^2 \right\rbrace

- bridge ordination (MDS) and \textit{generalised} semi-supervised learning
  - leverage $\sigma$ flexibility to vary duality principle

- data \textit{adaptivity} of semi-supervised learning
  - use to update dynamic graph $\leftrightarrow$ non-stationary time series
Graph diffusion

**Epidemic diffusion** (MOSAR): Apply standard tools...

▷ Relationship between virus spreading and graph structure: Can diffusion wavelets help?

▷ How to take into account / reflect dynamicity of graphs
Context and collaborations

- **Dante** (B. Girault, E. Fleury, ...)

- Other teams (e.g. Geodyn, Maestro...)

- International cooperations (e.g. EPFL)
Context and collaborations

- Dante (B. Girault, E. Fleury, ...)
- Institut des Systèmes Complexes
Context and collaborations

- **Dante** (B. Girault, E. Fleury, …)
- Institut des Systèmes Complexes
- Sisyphe (ENS Lyon, P. Borgnat)
Context and collaborations

- **Dante** (B. Girault, E. Fleury, ...)
- Institut des Systèmes Complexes
- Sisyphe (ENS Lyon, P. Borgnat)
- Other teams (e.g. Geodyn, Maestro, ...)

Follow-up: Dynamic graphs analysis
Context and collaborations

- **Dante** (B. Girault, E. Fleury, ...)
- Institut des Systèmes Complexes
- Sisyphe (ENS Lyon, P. Borgnat)
- Other teams (e.g. Geodyn, Maestro, ...)
- International cooperations (e.g. EPFL)
- ...