# Graph fibrations, graph isomorphism, and PageRank

**Paolo Boldi** Violetta Lonati
Massimo Santini Sebastiano Vigna

Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano

# Things related to PageRank

What do we speak of when we speak of PageRank?

- graphs
- (perturbed) Markov chains
- invariant distributions

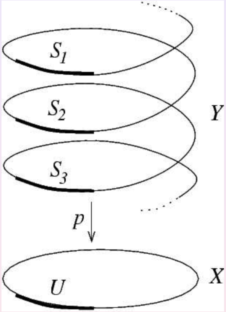
...and the other "usual suspects".

In this talk, some "unusual suspects" appear (for the first time on the screen)

- covering projections
- graph fibrations
- graph isomorphisms

# Covering projections in algebraic topology

- In algebraic topology, a *covering projection* is a continuous map that behaves *locally* like a homeomorphism:




Very roughly: it's a sort of local isomorphism.

# Covering projections in modern mathematics

- Every **graph** can be turned into a **topological space** by considering its geometric realization.
- This allows one to apply the definition of covering projections to graphs as well: in the case of graphs, the definition can actually be restated in purely combinatorial (and simple) form.
- In particular, covering projections became widely used in topological graph theory.

# From covering projections to fibrations

- Covering projections turn out to be too strong for many applications when *directed graphs* are involved.
- A weaker topological property, that of being a *fibration*, has been reformulated by Grothendieck for categories, and can be used naturally on graphs (seen as generators of categories).
- Grothendieck's notion of fibration boils down to a very simple one when applied to a graph.
- In fact, the community working on symbolic dynamics had independently defined fibrations and used them to classify shift systems and Markov chains up to measure-theoretic isomorphism [Ashley, Marcus & Tuncel, 1997].

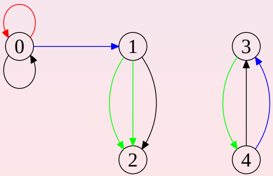
# My own personal relation with fibrations

- I first came in contact with fibrations when trying to solve (with Sebastiano Vigna) a problem in distributed computing:
  - given an anonymous (no ID's) message-passing asynchronous network...
  - ...under which conditions can the processors elect a leader.
- It turned out that this question can be answered completely using graph fibrations.
- We continued to use graph fibrations to solve various problems of distributed computability.
- Eventually, we collected all results on graph fibrations in a paper:

Paolo Boldi and Sebastiano Vigna. *Fibrations of graphs*. Discrete Math., 243:21-66, 2002

# A graph is a graph is a graph...

- In this case, generality makes things simpler.
- The word *graph* in this talk will always be used to mean
    - a set of nodes $N_G$ (usually: finite)
    - a set of arcs $A_G$ (usually: finite)
    - two maps $s_G : A_G \to N_G$ (source) and $t_G : A_G \to N_G$ (target)
    - a map $c_G : A_G \to C$ that assigns a colour to each arc.
- Loops are allowed; parallel arcs are allowed.
- When no parallel arcs exist, we say that the graph is *separated*.

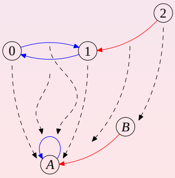


# Graph morphisms

- Given two graphs $G$ and $H$, a morphism $f : G \longrightarrow H$ maps nodes to nodes and arcs to arcs in such a way that sources, targets and colours are preserved.
- Formally:

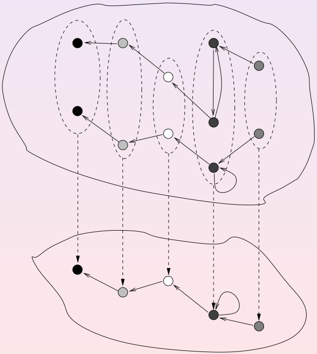
$$\begin{aligned} s_H(f(a)) &= f(s_G(a)) \\ t_H(f(a)) &= f(t_G(a)) \\ c_H(f(a)) &= c_G(a) \end{aligned}$$

for all arcs $a \in A_G$



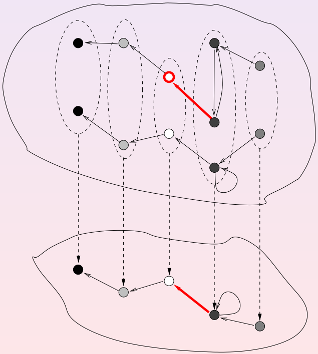
# Graph fibration

- A morphism $f : G \to H$ is a fibration if every arc of $H$ can be uniquely lifted, up to the choice of its target.
- Formally: for every arc $a \in A_H$ and every node $y \in N_G$ such that $f(y) = t(a)$, there is a unique arc $\widetilde{a}^y \in A_G$ such that $f(\widetilde{a}^y) = a$ and $t(\widetilde{a}^y) = y$.



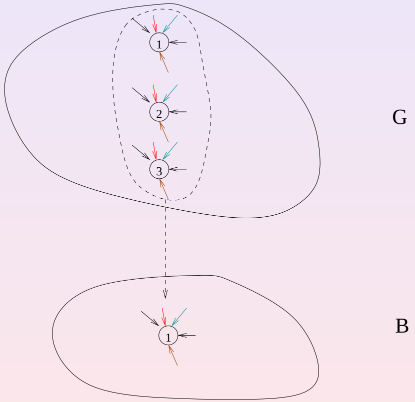
# Graph fibration

- A morphism $f : G \to H$ is a fibration if every arc of $H$ can be uniquely lifted, up to the choice of its target.
- Formally: for every arc $a \in A_H$ and every node $y \in N_G$ such that $f(y) = t(a)$, there is a unique arc $\widetilde{a}^y \in A_G$ such that $f(\widetilde{a}^y) = a$ and $t(\widetilde{a}^y) = y$.



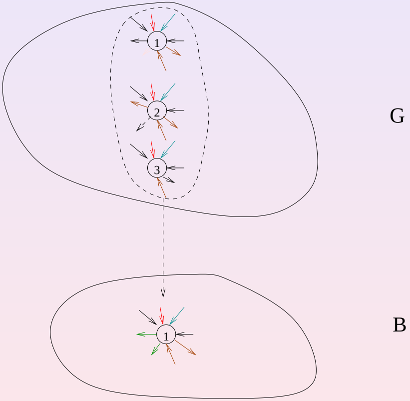
# A graph fibration is. . .

- A graph fibration is a local in-isomorphism.
- More explicitly: it is 1-1 on local in-neighborhoods



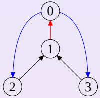
# A graph fibration is. . .

- A graph fibration is a local in-isomorphism.
- Nothing is required for out-neighborhoods!

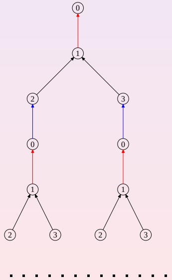


# A basic ingredient: universal total graph

- Let $G$ be a graph and $x$ a node of $G$

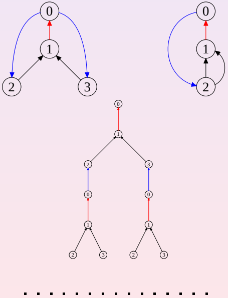



- The (usually infinite) tree of all paths ending in $x$ is called the *universal total graph* of $G$ at $x$, denoted by $\widetilde{G}^x$.



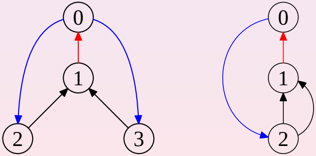
# Basic property of universal total graphs

- Let $G$ be a graph and $x$ a node of $G$
- Let $f : G \longrightarrow B$ be a fibration
- Then $\widetilde{G}^{x}$ and $\widetilde{B}^{f(x)}$ are isomorphic.
- Hence, in particular: two nodes of $G$ that are identified by some fibration must have isomorphic universal total graphs.



# Minimum base

- The converse is also true: if two nodes of $G$ have the same universal total graph, then they are identified by some fibration.
- More precisely, let $x \sim_G y$ whenever $\widetilde{G}^x$ and $\widetilde{G}^y$ are isomorphic.
- There is a graph $\widehat{G}$, whose nodes are the $\sim_G$-equivalence classes, such that $G$ is fibred over $\widehat{G}$.
- $\widehat{G}$ is called the *minimum base* of $G$.




$G$ $\widehat{G}$

# Markov chains and graphs

- A graph can be identified with the (transition matrix of a) Markov chain, provided that:
  - colors are non-negative real numbers (interpreted as transition probabilities)
  - for every node, the sum of the colors on outgoing arcs is 1:

$$\forall x \in N_G. \sum_{a:s_G(a)=x} c_G(a) = 1.$$

- Such graphs are called *stochastic*.
- The correspondence between stochastic graphs and row-stochastic matrices is 1-to-1 *for separated graphs*.

# Markov chains with restart

- Let $P$ be the transition matrix of a Markov chain; an *analytic perturbation* of $P$ [Schweitzer 1968] is

$$P(\varepsilon) ::= P + \varepsilon P_1 + \varepsilon^2 P_2 + \dots$$

  for small enough $\varepsilon$.

- We are going to consider a special case, where $P_2 = P_3 = \cdots = 0$ and $P_1$ has a special form: given a distribution $\mathbf{v}$ on the states:

$$\mathscr{R}(P, \mathbf{v}, \alpha) = \alpha P + (1 - \alpha)\mathbf{1}\mathbf{v}^T.$$

- Interpretation: at each step, with probability $\alpha$ we proceed as in $P$, with probability $1 - \alpha$ we "restart" from a state chosen according to $\mathbf{v}$; for this reason, $\mathscr{R}(P, \mathbf{v}, \alpha)$ is called a *Markov chain with restart*.

# PageRank as a special case

Standard PageRank can be seen as a special case of a Markov chain with restart:

$$\mathscr{R}(P, \mathbf{v}, \alpha) = \alpha P + (1 - \alpha)\mathbf{1}\mathbf{v}^T.$$

where:

- $P$ is the random-walk transition matrix defined on the graph: the probability to go from node $i$ to node $j$ in one step is

$$\begin{cases} 0 & \text{if there is no arc } i \to j \\ 1/d^+(i) & \text{if there is an arc } i \to j \text{ and } i \text{ has } d^+(i) \text{ outgoing arcs.} \end{cases}$$

- dangling nodes must be eliminated beforehand!

# PageRank: an example

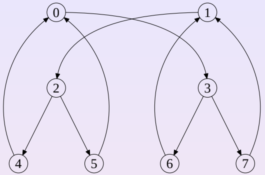



Figure: The graph

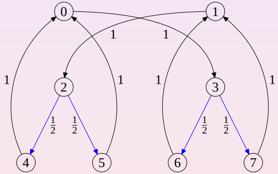



Figure: The corresponding Markov chain

# Markov chains with restart are unichain

**Theorem**

For every transition matrix $P$ and *every* preference vector $\mathbf{v}$:

- $\mathscr{R}(P, \mathbf{v}, \alpha)$ is unichain: all its essential (a.k.a. recurrent) states form a unique component;
- the essential states of $\mathscr{R}(P, \mathbf{v}, \alpha)$ are aperiodic.

As a consequence:

**Corollary**

$\mathscr{R}(P, \mathbf{v}, \alpha)$ has a unique invariant distribution $\mathbf{r}(P, \mathbf{v}, \alpha)$.

# Invariant distribution and limit behaviours

Some results about the invariant distribution $\mathbf{r}(P, \mathbf{v}, \alpha)$ of the Markov chain with restart $\mathscr{R}(P, \mathbf{v}, \alpha)$:

**Theorem**

- $$\mathbf{r}(P, \mathbf{v}, \alpha) = (1-\alpha)\mathbf{v}^T(I - \alpha P)^{-1}$$
- limit behaviour when $\alpha = 0$: $\mathbf{r}(P, \mathbf{v}, 0) = \mathbf{v}^T$
- limit behaviour when $\alpha \to 1$: $\lim_{\alpha \to 1^-} \mathbf{r}(P, \mathbf{v}, \alpha) = \mathbf{v}^T P^*$ where $P^*$ is the *Cesàro limit*

$$P^* = \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k.$$

# Power series associated to a graph

- Given an $\mathbf{R}^+$-coloured graph $G$, let $G^*(-, i)$ be the set of paths of $G$ ending in $i$; for every path $\pi$, let $c(\pi)$ be the *product* of the arc labels of $\pi$.
- For a distribution $\mathbf{v}$, define the following power series vector $\mathbf{s}(G, \mathbf{v}, \alpha)$

$$s_i(G, \mathbf{v}, \alpha) = (1 - \alpha) \sum_{t=0}^{\infty} \alpha^t \left( \sum_{\pi \in G^*(-,i), |\pi| = t} v_{s(\pi)} c(\pi) \right) .$$

- For a distribution $\mathbf{v}$, define the following power series vector $\mathbf{s}(G, \mathbf{v}, \alpha)$

$$s_i(G, \mathbf{v}, \alpha) = (1 - \alpha) \sum_{t=0}^{\infty} \alpha^t \boxed{\left( \sum_{\pi \in G^*(-,i), |\pi| = t} v_{s(\pi)} c(\pi) \right)} .$$

- The invariant distribution of a Markov chain with restart coincides with $\mathbf{s}(G, \mathbf{v}, \alpha)$; i.e., if $G$ is stochastic, then

$$\mathbf{s}(G, \mathbf{v}, \alpha) = \mathbf{r}(G, \mathbf{v}, \alpha).$$

# Power series and fibrations

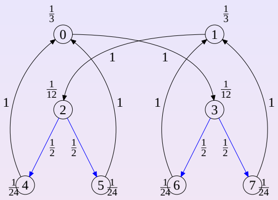
**Theorem**

Let $f : G \to B$ be a colour-preserving fibration and a distribution $\mathbf{v}$ on the nodes of $B$. Then:

$$\mathbf{s}(G, \mathbf{v}^f, \alpha) = \mathbf{s}(B, \mathbf{v}, \alpha)^f$$

... where $-^f$ means "copy along each fibre of $f$".

# An example



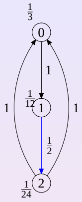




Figure: $\mathbf{s}(G, \mathbf{v}^f, \alpha) = \mathbf{s}(B, \mathbf{v}, \alpha)^f$

# Consequences

Implications of

$$\mathbf{s}(G, \mathbf{v}^f, \alpha) = \mathbf{s}(B, \mathbf{v}, \alpha)^f.$$

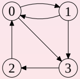
- Nodes of $G$ that are fibration equivalent *have the same PageRank (for all $\alpha$)* provided that the preference vector is fibrewise constant.
- Instead of computing $\mathbf{r}(G, \mathbf{v}^f, \alpha) = \mathbf{s}(G, \mathbf{v}^f, \alpha)$ one can compute $\mathbf{s}(B, \mathbf{v}, \alpha)$. This is advantageous! ($B$ can be much smaller!).
- Be careful: $B$ may not be stochastic, and $\mathbf{v}$ may not sum up to 1.
- Solution for the latter problems in the full paper.

# Markovian spectrally distinguishable graphs

- [Gori et al., 2005] proposed a polynomial isomorphism algorithm for the class of *Markovian spectrally distinguishable* graphs.
- A graph with $n$ nodes is *Markovian spectrally distinguishable* iff there are $n$ values $\alpha_0, \ldots, \alpha_{n-1}$ such that the PageRank vectors for these values form an invertible matrix.
- Since two nodes that are fibration equivalent have the same PageRank (for all $\alpha$'s), we have that:

  a Markovian spectrally distinguishable graph is fibration prime.

  (that is: it has no non-trivial fibrations)
- The converse is not true:



# Graph fibrations and graph isomorphism

- Graph isomorphism for fibration-prime graphs is polynomial.
- Hence, in particular, deciding isomorphism between Markovian spectrally distinguishable graphs can be done in polynomial time *with a completely combinatorial algorithm* (no PageRank computation required).
- Many practical algorithms for graph isomorphism exploit this fact.
- More precisely: they exploit the fact that nodes exchanged by an automorphism must have the same universal total graph.
- For example, McKay's famous `nauty` algorithm computes the minimum base, and then reasons on each fibre separately.
- But, how hard is it to compute the minimum base?

# Computing the minimum base

- The Cardon-Crochemore algorithm [Cardon and Crochemore, 1982] can be adapted to compute the minimum base (more precisely: to decide the $\sim_G$ relation) can be implemented with space occupancy $O(m + n)$ and time $O(m \log m \log n)$.
- Of course, this algorithm gives a necessary condition for Markovian distinguishability: if there are non-trivial equivalences, the graph is not Markovian spectrally distinguishable.
- For large graphs, $O(m + n)$ may be too much space: a different algorithm requires $O(n)$ space but with time $O(mn \log m \log n)$.

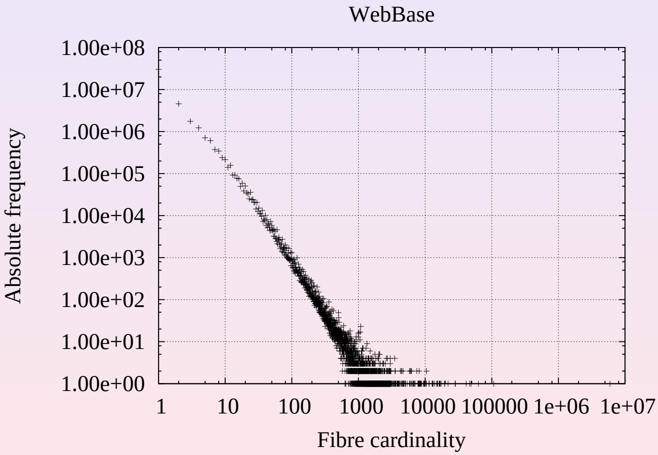
# Experimental results

We computed $\sim_G$ on some real Web graphs:

| Dataset | Number of nodes | Number of fibres | Avg. fibre size |
|---|---:|---:|---:|
| WebBase | 118,142,155 | 41,705,767 | 2.83 |
| `.it` | 41,291,594 | 15,245,587 | 2.71 |
| `.uk` | 39,459,925 | 14,154,663 | 2.79 |

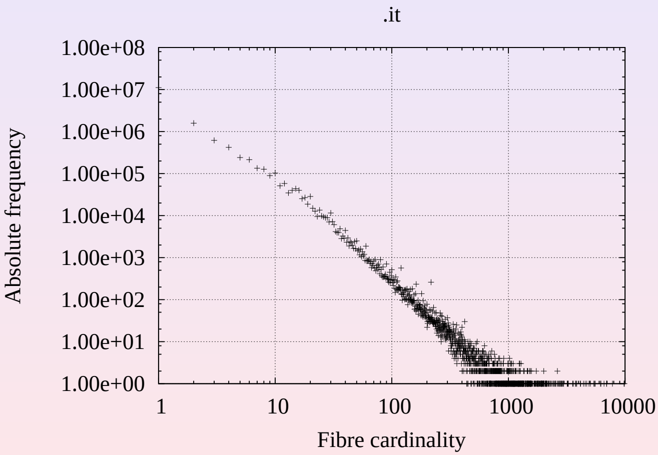
# Fibre cardinalities

Fibre cardinalities (in log/log scale):



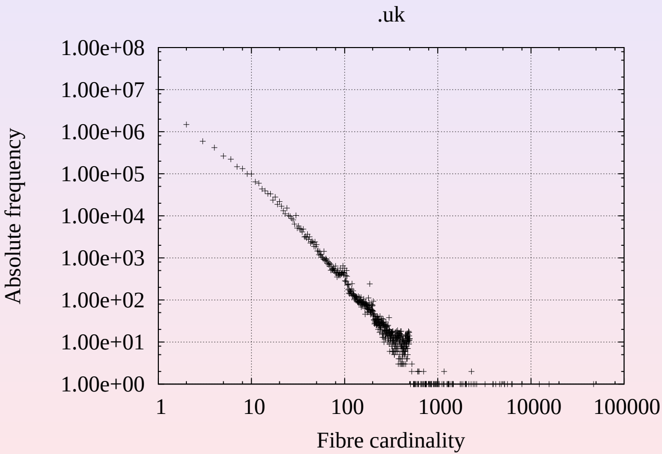
# Fibre cardinalities

Fibre cardinalities (in log/log scale):



# Fibre cardinalities

Fibre cardinalities (in log/log scale):



# Conclusions (and applications?)

- Computing $\sim_G$ gives a sufficient condition for two nodes to have the same PageRank (for all $\alpha$).
- No approximation! The algorithm is purely symbolic (combinatorial).
- PageRank can be computed on the minimum base — which is usually smaller.
- (But: computing the minimum base requires some time. . . )