Compression techniques
Graph

Paolo Boldi, DSI, Università degli studi di Milano

What do we mean by. . .

. . . "storing the web graph"?

- Having a data structure that allows you, for a given node, to know its successors.

What do we mean by. . .

. . . "storing the web graph"?

- Having a data structure that allows you, for a given node, to know its successors.
- Possibly: having a way to know which URL corresponds to a given node and vice versa.

What do we mean by. . .

. . . "storing the web graph"?

- Having a data structure that allows you, for a given node, to know its successors.
- Possibly: having a way to know which URL corresponds to a given node and vice versa.

We will study good data structures for the URL $\mapsto$ node map. The other one (less useful) is not covered here.

Information-theoretical lower bound

How much space do we need to store a graph with n nodes and m arcs?

Information-theoretical lower bound

How much space do we need to store a graph with n nodes and m arcs? Not less than

$$\log \binom{n^2}{m} \approx m \log \left(\frac{n^2}{m} \right) + O(m)$$

Information-theoretical lower bound

How much space do we need to store a graph with n nodes and m arcs? Not less than

$$\log \binom{n^2}{m} \approx m \log \left(\frac{n^2}{m} \right) + O(m)$$

under the hypothesis that $m = o(n^2)$

$$m \log \left(\frac{n}{d} \right) + O(m)$$

where $d = m/n$ is the average degree.

Information-theoretical lower bound

How much space do we need to store a graph with n nodes and m arcs? Not less than

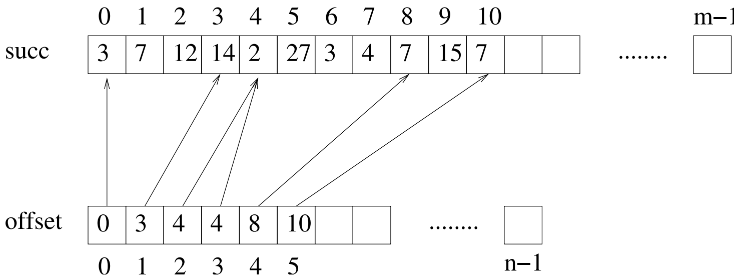
$$\log \binom{n^2}{m} \approx m \log \left(\frac{n^2}{m}\right) + O(m)$$

under the hypothesis that $m = o(n^2)$

$$m \log \left(\frac{n}{d}\right) + O(m)$$

where $d = m/n$ is the average degree.
This means about $\log(n/d) + O(1)$ bits per arc. But *social (web) graphs are NOT random graphs!*.

Naive representation



The offset vector tells, for each given node x, where the successor list of x starts from. Implicitly, it also gives the degree of each node.

Naive representation

How much space does this representation take?

- Successor array: m elements (arcs), each containing a node ($\log n$ bits); with 32 bits, we can store up to 4 billion nodes (half of it, if we don't have unsigned types)

Naive representation

How much space does this representation take?

- Successor array: m elements (arcs), each containing a node ($\log n$ bits); with 32 bits, we can store up to 4 billion nodes (half of it, if we don't have unsigned types)
- Offset array: n elements (nodes), each containing an index in the successor array ($\log m$ bits); with 32 bits, we can store up t 4 billion arcs.

Naive representation

How much space does this representation take?

- Successor array: m elements (arcs), each containing a node ($\log n$ bits); with 32 bits, we can store up to 4 billion nodes (half of it, if we don't have unsigned types)
- Offset array: n elements (nodes), each containing an index in the successor array ($\log m$ bits); with 32 bits, we can store up t 4 billion arcs.

All in all, $32(n+m)$ bits. If we assume $m = 8n$ (a very modest assumption on the outdegree), we need $288n$ bits, i.e., 288 bits/node, 36 bits/arc.
We show how to reduce this of an order of magnitude.

Idea

Use a variable-length representation for successors. Such a representation should obviously...

- be instantaneously decodable

Idea

Use a variable-length representation for successors. Such a representation should obviously...

- be instantaneously decodable
- minimize the expected bitlength.

Idea

Use a variable-length representation for successors. Such a representation should obviously. . .

- be instantaneously decodable
- minimize the expected bitlength.

What about the offset array?

Idea

Use a variable-length representation for successors. Such a representation should obviously...

- be instantaneously decodable
- minimize the expected bitlength.

What about the offset array?

- bit displacement vs. byte displacement (with alignment)

Idea

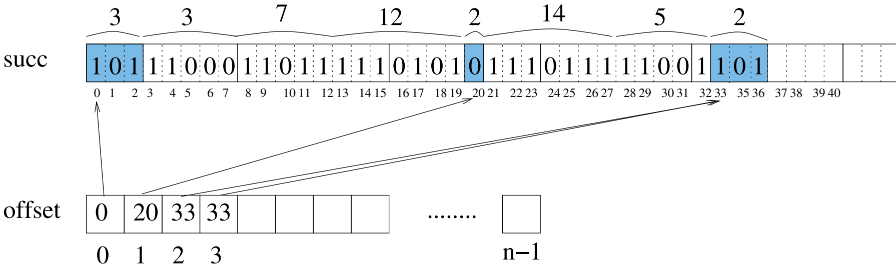
Use a variable-length representation for successors. Such a representation should obviously...

- be instantaneously decodable
- minimize the expected bitlength.

What about the offset array?

- bit displacement vs. byte displacement (with alignment)
- we have to keep an explicit representation of the node degrees (e.g., in the successor array, before every successsor list).

Variable-length representation



Node degrees (blue background), followed by successors. Each number is represented using an instantaneous code (possibly, different for degree and successors).

Instantaneous code

- An *instantaneous (binary) code* for the set S is a function $c : S \to \{0, 1\}^*$ such that, for all $x, y \in S$, if $c(x)$ is a prefix of $c(y)$, then $x = y$.

Instantaneous code

- An *instantaneous (binary) code* for the set S is a function $c : S \to \{0,1\}^*$ such that, for all $x, y \in S$, if $c(x)$ is a prefix of $c(y)$, then $x = y$.
- Let l_x be the length (in bits) of $c(x)$.

Instantaneous code

- An *instantaneous (binary) code* for the set S is a function $c : S \to \{0,1\}^*$ such that, for all $x, y \in S$, if $c(x)$ is a prefix of $c(y)$, then $x = y$.
- Let l_x be the length (in bits) of $c(x)$.
- Kraft-McMillan: there exists an instantaneous code with lengths l_x ($x \in S$) if and only if

$$\sum_{x \in S} 2^{-l_x} \leq 1.$$

Intended distribution

- If $p : S \to [0,1]$ is the probability distribution of the source, than the ideal code for S is such that (Shannon)

$$l_x = -\log p(x)$$

(all logarithms are in base 2).

Intended distribution

- If $p : S \to [0,1]$ is the probability distribution of the source, than the ideal code for S is such that (Shannon)

$$l_x = -\log p(x)$$

(all logarithms are in base 2).

- So a code with lengths l_x has *intended distribution*

$$p(x) = 2^{-l_x}.$$

Intended distribution

- If $p : S \to [0, 1]$ is the probability distribution of the source, than the ideal code for S is such that (Shannon)

$$l_x = -\log p(x)$$

  (all logarithms are in base 2).

- So a code with lengths l_x has *intended distribution*

$$p(x) = 2^{-l_x}.$$

- The choice of the code to use will be based on the expected distribution of the data.

Fixed-length coding

- If $S = \{1, 2, \ldots, N\}$, to represent an element of S it is sufficient to use $\lceil \log N \rceil$ bits.

Fixed-length coding

- If $S = \{1, 2, \ldots, N\}$, to represent an element of S it is sufficient to use $\lceil \log N \rceil$ bits.
- The fixed-length representation for S uses exactly that number of bits for every element (and represents x using the standard binary coding of $x - 1$ on $\lceil \log N \rceil$ bits).

Fixed-length coding

- If $S = \{1, 2, \ldots, N\}$, to represent an element of S it is sufficient to use $\lceil \log N \rceil$ bits.
- The fixed-length representation for S uses exactly that number of bits for every element (and represents x using the standard binary coding of $x - 1$ on $\lceil \log N \rceil$ bits).
- Intended distribution:

$$p(x) = 2^{-\lceil \log N \rceil} \quad \text{uniform distribution.}$$

Unary coding

- If $S = \mathbf{N}$, one can represent $x \in S$ writing x zeroes followed by a one.

Unary coding

- If $S = \mathbf{N}$, one can represent $x \in S$ writing x zeroes followed by a one.
- So $l_x = x + 1$, and the intended distribution is

$$p(x) = 2^{-x-1} \quad \text{geometric distribution of ratio } 1/2.$$

Unary coding

- If $S = \mathbf{N}$, one can represent $x \in S$ writing x zeroes followed by a one.
- So $l_x = x + 1$, and the intended distribution is

$$p(x) = 2^{-x-1} \quad \text{geometric distribution of ratio } 1/2.$$

0	1
1	01
2	001
3	0001
4	00001

A more general viewpoint

Unary coding can be seen as a special case of a more general kind of coding for **N**. Suppose you group **N** into *slots*: every slot is made by consecutive integers; let

$$V = \langle s_1, s_2, s_3, \dots \rangle$$

be the slot sizes (in the unary case $s_1 = s_2 = \dots = 1$).

A more general viewpoint

Unary coding can be seen as a special case of a more general kind of coding for **N**. Suppose you group **N** into *slots*: every slot is made by consecutive integers; let

$$V = \langle s_1, s_2, s_3, \dots \rangle$$

be the slot sizes (in the unary case $s_1 = s_2 = \cdots = 1$).
Then, to represent $x \in \mathbf{N}$ one can

A more general viewpoint

Unary coding can be seen as a special case of a more general kind of coding for **N**. Suppose you group **N** into *slots*: every slot is made by consecutive integers; let

$$V = \langle s_1, s_2, s_3, \dots \rangle$$

be the slot sizes (in the unary case $s_1 = s_2 = \cdots = 1$).
Then, to represent $x \in \mathbf{N}$ one can

- encode *in unary* the index i of the slot containing x;

A more general viewpoint

Unary coding can be seen as a special case of a more general kind of coding for **N**. Suppose you group **N** into *slots*: every slot is made by consecutive integers; let

$$V = \langle s_1, s_2, s_3, \dots \rangle$$

be the slot sizes (in the unary case $s_1 = s_2 = \cdots = 1$).
Then, to represent $x \in \mathbf{N}$ one can

- encode *in unary* the index i of the slot containing x;
- encode *in binary* the offset of x within its slot (using $\lceil \log s_i \rceil$ bits).

Golomb coding

Golomb coding with modulus b is obtained choosing

$$V = \langle b, b, b, \dots \rangle.$$

To represent $x \in \mathbf{N}$ you need to specify the slot where x falls (that is, $\lfloor x/b \rfloor$) in unary, and then represent the offset using $\lceil \log b \rceil$ bits (or $\lfloor \log b \rfloor$ bits).

Golomb coding

Golomb coding with modulus b is obtained choosing

$$V = \langle b, b, b, \dots \rangle.$$

To represent $x \in \mathbf{N}$ you need to specify the slot where x falls (that is, $\lfloor x/b \rfloor$) in unary, and then represent the offset using $\lceil \log b \rceil$ bits (or $\lfloor \log b \rfloor$ bits).
So

$$l_x = \left\lfloor \frac{x}{b} \right\rfloor + \lceil \log b \rceil.$$

The intended distribution is

$$p(x) = 2^{-l_x} \propto (2^{1/b})^{-x} \quad \text{geometric distribution of ratio } 1/\sqrt[b]{2}.$$

More precisely. . .

A finer analysis shows that Golomb coding is optimal (=Huffman) for a geometric distribution of ratio p, provided that b is chosen as

$$b = \left\lceil \frac{\log(2-p)}{-\log(1-p)} \right\rceil .$$

More precisely. . .

A finer analysis shows that Golomb coding is optimal (=Huffman) for a geometric distribution of ratio p, provided that b is chosen as

$$b = \left\lceil \frac{\log(2-p)}{-\log(1-p)} \right\rceil .$$

0	**1**0
1	**1**10
2	**1**11
3	**01**0
4	**01**10
5	**01**11
6	**001**0

Elias' γ

Elias' γ *coding* of $x \in \mathbf{N}^+$ is obtained by representing x in binary preceded by a unary representation of its length (minus one).

Elias' γ

Elias' γ *coding* of $x \in \mathbf{N}^+$ is obtained by representing x in binary preceded by a unary representation of its length (minus one). More precisely, to represent x we write in unary $\lfloor \log x \rfloor$ and then in binary $x - 2^{\lceil \log x \rceil}$ (on $\lfloor \log x \rfloor$ bits).

Elias' γ

Elias' γ *coding* of $x \in \mathbf{N}^+$ is obtained by representing x in binary preceded by a unary representation of its length (minus one). More precisely, to represent x we write in unary $\lfloor \log x \rfloor$ and then in binary $x - 2^{\lceil \log x \rceil}$ (on $\lfloor \log x \rfloor$ bits). So

$$l_x = 1 + 2\lfloor \log x \rfloor \quad \Longrightarrow \quad p(x) \propto \frac{1}{2x^2} \text{(Zipf of exponent 2)}$$

Elias' γ

Elias' γ *coding* of $x \in \mathbf{N}^+$ is obtained by representing x in binary preceded by a unary representation of its length (minus one). More precisely, to represent x we write in unary $\lfloor \log x \rfloor$ and then in binary $x - 2^{\lceil \log x \rceil}$ (on $\lfloor \log x \rfloor$ bits). So

$$l_x = 1 + 2\lfloor \log x \rfloor \quad \Longrightarrow \quad p(x) \propto \frac{1}{2x^2} (\text{Zipf of exponent } 2)$$

1	**1**
2	**01**0
3	**01**1
4	**001**00
5	**001**01

Elias' δ

Elias' δ *coding* of $x \in \mathbf{N}^+$ is obtained by representing x in binary preceded by a representation of its length in γ.

Elias' δ

Elias' δ *coding* of $x \in \mathbf{N}^+$ is obtained by representing x in binary preceded by a representation of its length in γ.
So

$$l_x = 1 + 2\lfloor \log\log x \rfloor + \lfloor \log x \rfloor \qquad \Longrightarrow \; p(x) \propto \frac{1}{2x(\log x)^2}$$

Elias' δ

Elias' δ *coding* of $x \in \mathbf{N}^+$ is obtained by representing x in binary preceded by a representation of its length in γ.
So

$$l_x = 1 + 2\lfloor \log\log x \rfloor + \lfloor \log x \rfloor \quad \Longrightarrow \quad p(x) \propto \frac{1}{2x(\log x)^2}$$

1	**1**
2	**010**0
3	**010**1
4	**011**00
5	**011**01
6	**00100**000
7	**00100**001

An alternative way. . .

. . . to think of γ coding is that x is represented using its usual binary representation (except for the initial "1", which is omitted), with every bit "coming with" a continuation bit, that tells whether the representation continues or whether it stops there.
For example (up to bit permutation) γ coding of 724 (in binary: 1011010100) is

0	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	0	0

k-bit-variable coding

What happens if we group digits k by k?

0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0

0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0

0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0

1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0

k-bit-variable coding (cont'd)

For x, we use $\lceil \log(x)/k \rceil$ bits for the unary part, and the same number of bits multiplied by k for the binary part.

k-bit-variable coding (cont'd)

For x, we use $\lceil \log(x)/k \rceil$ bits for the unary part, and the same number of bits multiplied by k for the binary part.
So

$$l_x = (k+1)(\lceil \log(x)/k \rceil) \quad \Longrightarrow \; p(x) \propto x^{-(k+1)/k} (\text{Zipf } (k+1)/k)$$

k-bit-variable coding (cont'd)

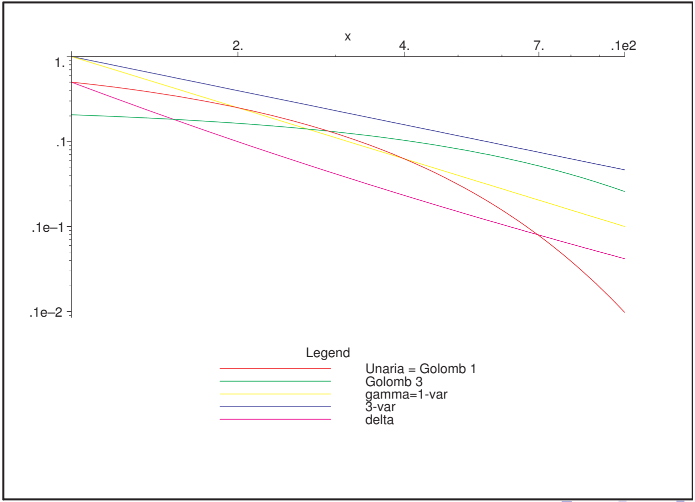
For x, we use $\lceil \log(x)/k \rceil$ bits for the unary part, and the same number of bits multiplied by k for the binary part.
So

$$l_x = (k+1)(\lceil \log(x)/k \rceil) \implies p(x) \propto x^{-(k+1)/k} (\text{Zipf } (k+1)/k)$$

A more efficient variant: the ζ_k codes (for Zipf $1 \to 2$).

	$\gamma = \zeta_1$	ζ_2	ζ_3	ζ_4
1	1	10	100	1000
2	010	110	1010	10010
3	011	111	1011	10011
4	00100	01000	1100	10100
5	00101	01001	1101	10101
6	00110	01010	1110	10110
7	00111	01011	1111	10111
8	0001000	011000	0100000	11000

Comparing codings



Coding techniques. . .

. . . alone do not improve on compression: we have first to guarantee that the data we represent have a distribution close to the intended one (depending on the coding we are going to use). In particular, they have to enjoy a monotonic distribution (smaller values are more probable than larger ones).

Coding techniques. . .

. . . alone do not improve on compression: we have first to guarantee that the data we represent have a distribution close to the intended one (depending on the coding we are going to use). In particular, they have to enjoy a monotonic distribution (smaller values are more probable than larger ones).

- BTW: some codings (e.g., Elias γ and δ) are universal: for whatever monotonic distribution, they guarantee an expected length that is only within a constant factor of the optimal one.

Coding techniques. . .

. . . alone do not improve on compression: we have first to guarantee that the data we represent have a distribution close to the intended one (depending on the coding we are going to use). In particular, they have to enjoy a monotonic distribution (smaller values are more probable than larger ones).

- BTW: some codings (e.g., Elias γ and δ) are universal: for whatever monotonic distribution, they guarantee an expected length that is only within a constant factor of the optimal one.
- Degrees are distributed like a Zipf of exponent ≈ 2.7: they can be safely encoded using γ.

Coding techniques...

...alone do not improve on compression: we have first to guarantee that the data we represent have a distribution close to the intended one (depending on the coding we are going to use). In particular, they have to enjoy a monotonic distribution (smaller values are more probable than larger ones).

- BTW: some codings (e.g., Elias γ and δ) are universal: for whatever monotonic distribution, they guarantee an expected length that is only within a constant factor of the optimal one.
- Degrees are distributed like a Zipf of exponent ≈ 2.7: they can be safely encoded using γ.
- What about successors? Let us assume that successors of x are $y_1, \ldots, y_k$: how should we encode $y_1, \ldots, y_k$?

Locality

In general, we cannot say much about their distribution, unless we make some assumption on the way in which nodes are ordered.

Locality

In general, we cannot say much about their distribution, unless we make some assumption on the way in which nodes are ordered.

- Many hypertextual links contained in a web page are *navigational* ("home", "next", "up"...). If we compare the URL they refer to with that of the page containing them, they share a long common prefix. This property is known as *locality* and it was first observed by the authors of the Connectivity Server.

Locality

In general, we cannot say much about their distribution, unless we make some assumption on the way in which nodes are ordered.

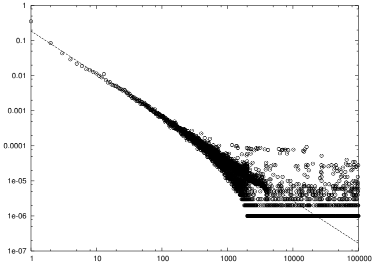
- Many hypertextual links contained in a web page are *navigational* ("home", "next", "up"...). If we compare the URL they refer to with that of the page containing them, they share a long common prefix. This property is known as *locality* and it was first observed by the authors of the Connectivity Server.
- To exploit this property, assume that URLs are ordered lexicographically (that is, node 0 is the first URL in lexicographic order, etc.). Then, if $x \to y$ is an arc, most of the times $|x - y|$ will be "small".

Exploiting locality

If x has successors $y_1 < y_2 < \cdots < y_k$, we represent its successor list though the gaps (*differentiation*):

$$y_1 - x, y_2 - y_1 - 1, \ldots, y_k - y_{k-1} - 1$$

(only the first value can be negative: wa make it into a natural number. . .). How are such differences distributed?



Zipf with exponent 1.2 $\Longrightarrow$ we use ζ_3.

Similarity

URLs close to each other (in lexicographic order) have similar successor sets: this fact (known as *similarity*) was exploited for the first time in the Link database.

Similarity

URLs close to each other (in lexicographic order) have similar successor sets: this fact (known as *similarity*) was exploited for the first time in the Link database.
We may encode the successor list of x as follows:

Similarity

URLs close to each other (in lexicographic order) have similar successor sets: this fact (known as *similarity*) was exploited for the first time in the Link database.

We may encode the successor list of x as follows:

- we write the differences with respect to the successor list of some previous node $x - r$ (called the *reference node*)

Similarity

URLs close to each other (in lexicographic order) have similar successor sets: this fact (known as *similarity*) was exploited for the first time in the Link database.

We may encode the successor list of x as follows:

- we write the differences with respect to the successor list of some previous node $x - r$ (called the *reference node*)
- we explicitly encode (as before) only the successors of x that were not successors of $x - r$.

Similarity (cont'd)

More explicitly, the successor list of x is encoded as (*referencing*):

- an intger r (reference): if $r > 0$, the list is described by difference with respect to the successor list of $x - r$; in this case, we write a bitvector (of length equal to $d^+(x-r)$) discriminating the elements in $N^+(x-r) \cap N^+(x)$ from the ones in $N^+(x-r) \setminus N^+(x)$

Similarity (cont'd)

More explicitly, the successor list of x is encoded as (*referencing*):

- an intger r (reference): if $r > 0$, the list is described by difference with respect to the successor list of $x - r$; in this case, we write a bitvector (of length equal to $d^+(x - r)$) discriminating the elements in $N^+(x - r) \cap N^+(x)$ from the ones in $N^+(x - r) \setminus N^+(x)$
- an explicit list of *extra nodes*, containing the elements of $N^+(x) \setminus N^+(x - r)$ (or the whole $N^+(x)$, if $r = 0$), encoded as explained before.

Referencing example

Node	Outdegree	Successors
...	...	...
15	11	13, 15, 16, 17, 18, 19, 23, 24, 203, 315, 1034
16	10	15, 16, 17, 22, 23, 24, 315, 316, 317, 3041
17	0	
18	5	13, 15, 16, 17, 50
...	...	...

Node	Outd.	Ref.	Copy list	Extra nodes
...	...	...	...	...
15	11	0		13, 15, 16, 17, 18, 19, 23, 24, 203, 315, 1034
16	10	1	01110011010	22, 316, 317, 3041
17	0			
18	5	3	11110000000	50
...	...	...	...	...

Blocks (*differential compression*)

Instead of using a bitvector, we use run-length encoding, telling the length of successive runs (blocks) of “0” and “1”:

Node	Outd.	Ref.	Copy list	Extra nodes
...	...	...	...	...
15	11	0		13, 15, 16, 17, 18, 19, 23, 24, 203, 315, 1034
16	10	1	01110011010	22, 316, 317, 3041
17	0			
18	5	3	11110000000	50
...	...	...	...	...

Blocks (*differential compression*)

Instead of using a bitvector, we use run-length encoding, telling the length of successive runs (blocks) of “0” and “1”:

Node	Outd.	Ref.	Copy list	Extra nodes
...	...	...	...	...
15	11	0		13, 15, 16, 17, 18, 19, 23, 24, 203, 315, 1034
16	10	1	01110011010	22, 316, 317, 3041
17	0			
18	5	3	11110000000	50
...	...	...	...	...

Node	Outd.	Ref.	# blocks	Copy blocks	Extra nodes
...	...	...	...	...	...
15	11	0			13, 15, 16, 17, 18, 19, 23, ...
16	10	1	7	0, 0, 2, 1, 1, 0, 0	22, 316, ...
17	0				
18	5	3	1	4	50
...	...	...	...	...	...

Consecutivity

Among the extra nodes, many happen to sport the *consecutivity* property: they appear in clusters of consecutive integers. This phenomenon, observed empirically, have some possible explanations:

Consecutivity

Among the extra nodes, many happen to sport the *consecutivity* property: they appear in clusters of consecutive integers. This phenomenon, observed empirically, have some possible explanations:

- most pages contain groups of navigational links that correspond to a certain hierarchical level of the website, and are often consecutive to one another;

Consecutivity

Among the extra nodes, many happen to sport the *consecutivity* property: they appear in clusters of consecutive integers. This phenomenon, observed empirically, have some possible explanations:

- most pages contain groups of navigational links that correspond to a certain hierarchical level of the website, and are often consecutive to one another;
- in the transpose graph, moreover, consecutivity is the dual of similarity with reference 1: when there is a cluster of consecutive pages with many similar links, in the transpose there are intervals of consecutive outgoing links.

Consecutivity (cont'd)

To exploit consecutivity, we use a special representation for the extra node list called *intervalization*, that is:

- sufficiently long (say $\geq T$) intervals of consecutive integers are represented by their left extreme and their length minus T;
- other extra nodes, if any, are called *residual nodes* and are represented alone.

Intervalization example

Node	Outd.	Ref.	# blocks	Copy blocks	Extra nodes
...	...	...	...	...	...
15	11	0			13, 15, 16, 17, 18, 19, 23, ...
16	10	1	7	0, 0, 2, 1, 1, 0, 0	22, 316, ...
17	0				
18	5	3	1	4	50
...	...	...	...	...	...

Intervalization example

Node	Outd.	Ref.	# blocks	Copy blocks	Extra nodes
...	...	...	...	...	...
15	11	0			13, 15, 16, 17, 18, 19, 23, ...
16	10	1	7	0, 0, 2, 1, 1, 0, 0	22, 316, ...
17	0				
18	5	3	1	4	50
...	...	...	...	...	...

Node	Outd.	Ref.	# bl.	Copy bl.s	# int.	Lft extr.	Lth	Residuals
...	...	...	...	...	...	...	...	...
15	11	0			2	15,...	4,...	13, 23 ...
16	10	1	7	0, 0, ...	1	316	1	22, 3041
17	0							
18	5	3	1	4	0			50
...	...	...	...	...	...	...	...	...

Reference window

When the reference node is chosen, how far back in the “past” are we allowed to go?

Reference window

When the reference node is chosen, how far back in the “past” are we allowed to go? We need to keep track of a window of the last W successor lists. The choice of W is critical:

Reference window

When the reference node is chosen, how far back in the “past” are we allowed to go? We need to keep track of a window of the last W successor lists. The choice of W is critical:

- a large W guarantees better compression, but increases compression time and space

Reference window

When the reference node is chosen, how far back in the “past” are we allowed to go? We need to keep track of a window of the last W successor lists. The choice of W is critical:

- a large W guarantees better compression, but increases compression time and space
- after $W = 7$ there is no significant improvement in compression.

Reference window

When the reference node is chosen, how far back in the "past" are we allowed to go? We need to keep track of a window of the last W successor lists. The choice of W is critical:

- a large W guarantees better compression, but increases compression time and space
- after $W = 7$ there is no significant improvement in compression.

The choice of W does not impact on decompression time.

Reference chain length

Referencing involves recursion: to decode the successor list of x, we need first to decompress the successor list of $x - r$, etc. This chain is called the *reference chain* of x: decompression speed depends on the length of such chains.

Reference chain length

Referencing involves recursion: to decode the successor list of x, we need first to decompress the successor list of $x - r$, etc. This chain is called the *reference chain* of x: decompression speed depends on the length of such chains.
During compression, it is possible to limit their length keeping into account of how long is the reference chain for every node in the window and avoiding to use nodes whose reference chain is already of a given maximum length R.

Reference chain length

Referencing involves recursion: to decode the successor list of x, we need first to decompress the successor list of $x - r$, etc. This chain is called the *reference chain* of x: decompression speed depends on the length of such chains.

During compression, it is possible to limit their length keeping into account of how long is the reference chain for every node in the window and avoiding to use nodes whose reference chain is already of a given maximum length R.

The choice of R influences the compression ratio (with $R = \infty$ giving the best possible compression) but also on decompression speed ($R = \infty$ may produce access time that can be two orders of magnitude larger than $R = 1$ — it may even produce stack overflows).

From web graphs to social networks

The basic property we have been exploiting so far is that *nodes are numbered according to the lexicographic ordering of URLs*. Is it possible to adapt / extend this idea to non-web graphs, e.g., to social networks?

From web graphs to social networks

The basic property we have been exploiting so far is that *nodes are numbered according to the lexicographic ordering of URLs*. Is it possible to adapt / extend this idea to non-web graphs, e.g., to social networks?

- What we want is an ordering of the nodes that is compression friendly

From web graphs to social networks

The basic property we have been exploiting so far is that *nodes are numbered according to the lexicographic ordering of URLs*. Is it possible to adapt / extend this idea to non-web graphs, e.g., to social networks?

- What we want is an ordering of the nodes that is compression friendly
- In particular, we want that most arcs are between nodes that are very close (as numbers) to each other.

Orderings and communities

Social networks have no natural ordering such as “lexicographic by URL”. However many statistics suggest that social networks are clustered.

Orderings and communities

Social networks have no natural ordering such as “lexicographic by URL”. However many statistics suggest that social networks are clustered.
Goal: unravel the clustered structure inside social networks, search for an ordering that run through clusters and use it to compress the graph much better than the theoretical lower bound.

Orderings and communities

Social networks have no natural ordering such as “lexicographic by URL”. However many statistics suggest that social networks are clustered.
Goal: unravel the clustered structure inside social networks, search for an ordering that run through clusters and use it to compress the graph much better than the theoretical lower bound.
Constraints:

Orderings and communities

Social networks have no natural ordering such as “lexicographic by URL”. However many statistics suggest that social networks are clustered.
Goal: unravel the clustered structure inside social networks, search for an ordering that run through clusters and use it to compress the graph much better than the theoretical lower bound.
Constraints:

1. very few clustering techniques scale up to very large graphs

Orderings and communities

Social networks have no natural ordering such as "lexicographic by URL". However many statistics suggest that social networks are clustered.
Goal: unravel the clustered structure inside social networks, search for an ordering that run through clusters and use it to compress the graph much better than the theoretical lower bound.
Constraints:

1. very few clustering techniques scale up to very large graphs
2. we do not posses any prior information on the number of clusters

Orderings and communities

Social networks have no natural ordering such as "lexicographic by URL". However many statistics suggest that social networks are clustered.

Goal: unravel the clustered structure inside social networks, search for an ordering that run through clusters and use it to compress the graph much better than the theoretical lower bound.

Constraints:

1. very few clustering techniques scale up to very large graphs
2. we do not posses any prior information on the number of clusters
3. cluster sizes are going to be very unbalanced

Orderings and communities (cont'd)

Orderings and communities (cont'd)

- You can obtain an ordering from a clustering just sorting by cluster label

Orderings and communities (cont'd)

- You can obtain an ordering from a clustering just sorting by cluster label
- Different clustering algorithms yield different and incomparable orderings

Orderings and communities (cont'd)

- You can obtain an ordering from a clustering just sorting by cluster label
- Different clustering algorithms yield different and incomparable orderings
- Main idea:

Orderings and communities (cont'd)

- You can obtain an ordering from a clustering just sorting by cluster label
- Different clustering algorithms yield different and incomparable orderings
- Main idea:
  - Run a clustering algorithm A

Orderings and communities (cont'd)

- You can obtain an ordering from a clustering just sorting by cluster label
- Different clustering algorithms yield different and incomparable orderings
- Main idea:
  - Run a clustering algorithm A
  - Renumber nodes sorting by A's labels, breaking ties using the node numbers (i.e., sort stably by A's labels)

Orderings and communities (cont'd)

- You can obtain an ordering from a clustering just sorting by cluster label
- Different clustering algorithms yield different and incomparable orderings
- Main idea:
  - Run a clustering algorithm A
  - Renumber nodes sorting by A's labels, breaking ties using the node numbers (i.e., sort stably by A's labels)
  - Iterate with another clustering algorithm

Label Propagation Algorithm (LPA)

LPA are a class of clustering algorithm that work as follows:

Label Propagation Algorithm (LPA)

LPA are a class of clustering algorithm that work as follows:

- Every node adopts the label that is most common among its neighbors. . .

Label Propagation Algorithm (LPA)

LPA are a class of clustering algorithm that work as follows:

- Every node adopts the label that is most common among its neighbors. . .
- . . . with an adjustment depending on the overall popularity of the label

Label Propagation Algoritm (LPA)

Require: G a graph, γ a density parameter

$\pi \leftarrow$ a random permutation of G's nodes
for all x: $\lambda(x) \leftarrow x$, $v(x) \leftarrow 1$
while (some stopping criterion) **do**
  for $i = 0, 1, \ldots, n-1$ **do**
    for every label ℓ, $k_\ell \leftarrow |\lambda^{-1}(\ell) \cap N_G(\pi(i))|$
    $\hat{\ell} \leftarrow \mathrm{argmax}_\ell[k_\ell - \gamma(v(\ell) - k_\ell)]$
    decrement $v(\lambda(\pi(i)))$
    $\lambda(\pi(i)) \leftarrow \hat{\ell}$
    increment $v(\lambda(\pi(i)))$
  end for
end while

Here $v(\ell)$ is the number of nodes currently labelled by ℓ, so $v(\ell) - k_\ell$ is the popularity of label ℓ outside of the current neighborhood.

Layered Label Propagation Algoritm (LLPA)

Layered Label Propagation Algoritm (LLPA)

- Repeatedly run LPA with different values of γ

Layered Label Propagation Algoritm (LLPA)

- Repeatedly run LPA with different values of γ
- Renumber nodes sorting by stably by the new labels

Name	LLP		BFS	Shingle		Natural		Random	
Amazon	9.16	(-30%)	12.96	14.43	(+11%)	16.92	(+30%)	23.62	(+82%)
DBLP	6.88	(-23%)	8.91	11.42	(+28%)	11.36	(+27%)	22.07	(+147%)
Enron	6.51	(-24%)	8.54	9.87	(+15%)	13.43	(+57%)	14.02	(+64%)
Hollywood	5.14	(-35%)	7.81	6.72	(-14%)	15.20	(+94%)	16.23	(+107%)
LiveJournal	10.90	(-28%)	15.1	15.77	(+4%)	14.35	(-5%)	23.50	(+55%)
Flickr	8.89	(-22%)	11.26	10.22	(-10%)	13.87	(+23%)	14.49	(+28%)
indochina (hosts)	5.53	(-17%)	6.63	7.16	(+7%)	9.26	(+39%)	10.59	(+59%)
uk (hosts)	6.26	(-18%)	7.62	8.12	(+6%)	10.81	(+41%)	15.58	(+104%)
eu	3.90	(-21%)	4.93	6.86	(+39%)	5.24	(+6%)	19.89	(+303%)
in	2.46	(-30%)	3.51	4.79	(+36%)	2.99	(-15%)	21.15	(+502%)
indochina	1.71	(-26%)	2.31	3.59	(+55%)	2.19	(-6%)	21.46	(+829%)
it	2.10	(-28%)	2.89	4.39	(+51%)	2.83	(-3%)	26.40	(+813%)
uk	1.91	(-33%)	2.84	4.09	(+44%)	2.75	(-4%)	27.55	(+870%)
altavista-nd	5.22	(-11%)	5.85	8.12	(+38%)	8.37	(+43%)	34.76	(+494%)