Extending de Bruijn sequences to larger alphabets

Verónica Becher

Universidad de Buenos Aires & CONICET

Joint work with Lucas Cortés

STIC AMSUD RAPA2 Project - December 11, 2020

de Bruijn sequence

A circular sequence is the equivalence class of a sequence under rotations.
Example: $[abc]$ is the circular sequence formed by the rotations of abc.

A de Bruijn sequence of order n over a k-symbol alphabet is a circular sequence of length k^n in which every length-n sequence occurs once.

Example: $[0011]$ is de Bruijn of order 2 over the alphabet $\{0, 1\}$.

The Problem

Fix an alphabet.

Given a de Bruijn sequence of order n over this alphabet,
Can we insert a new symbol (and possibly old ones) to obtain
a de Bruijn sequence of order n over the enlarged alphabet?

The Problem

Fix an alphabet.

Given a de Bruijn sequence of order n over this alphabet,
Can we insert a new symbol (and possibly old ones) to obtain
a de Bruijn sequence of order n over the enlarged alphabet?

The Problem

Fix an alphabet.

Given a de Bruijn sequence of order n over this alphabet,
Can we insert a new symbol (and possibly old ones) to obtain
a de Bruijn sequence of order n over the enlarged alphabet?

We do not want long runs without the new symbol.

The Problem

Fix an alphabet.

Given a de Bruijn sequence of order n over this alphabet,
Can we insert a new symbol (and possibly old ones) to obtain
a de Bruijn sequence of order n over the enlarged alphabet?

We do not want long runs without the new symbol.

For example, given this Bruijn sequence of order 3 over alphabet $\{0, 1\}$,

$$[11000101]$$

the following de Bruijn sequence of order 3 over alphabet $\{0, 1, 2\}$,

$$[1\ 2221211\ 100\ 220200\ 01\ 201021\ 01]$$

Subsequence

A subsequence of a sequence $a_1 a_2 \ldots a_n$ is a sequence $b_1 b_2 \ldots b_m$ defined by $b_i = a_{n_i}$ for $i = 1, 2, \ldots, m$, where $n_1 \leq n_2 \leq \ldots \leq n_m$.
The same applies to circular sequences, assuming any starting position.
Example: $[123]$, $[246]$ and [5612] are subsequences of $[123456]$.

Obviously

The extension problem is easy if you do not ask that there are no long runs without the new symbol.

Obviously

The extension problem is easy if you do not ask that there are no long runs without the new symbol.

Easy because

1. Each de Bruijn sequence of order $n+1$ is an Eulerian cycle in the de Bruijn graph of rder $n-1$, $G_{k,n}$.
2. $G_{k,n}$ is a proper subgraph of $G_{k+1,n}$
3. $G_{k+1,n} \setminus Gk, n$ is Eulerian
4. The extension problem is solved by Hierholzer's algorithm for joining cycles together to form Eulerian cycle.

No long runs without the new symbol

Given de Bruijn sequence of order n, consider the insertion of symbols (all the old symbols and one new).

In between every two successive occurrences of the new symbol

1. Fewer than n symbols?

No long runs without the new symbol

Given de Bruijn sequence of order n, consider the insertion of symbols (all the old symbols and one new).

In between every two successive occurrences of the new symbol

1. Fewer than n symbols?
   No, It would be impossible to accommodate all words of length n lacking the new symbol.
2. Exactly n symbols?

No long runs without the new symbol

Given de Bruijn sequence of order n, consider the insertion of symbols (all the old symbols and one new).

In between every two successive occurrences of the new symbol

1. Fewer than n symbols?
   No, It would be impossible to accommodate all words of length n lacking the new symbol.
2. Exactly n symbols?
   No, it would be impossible because to accommodate all words of length n lacking the new symbol we would need $(n+1)k^n$ symbols which exceeds $(k+1)^n$.
3. How many more than n symbols?

Theorem

For any de Bruijn sequence v of order n over a k-symbol alphabet there is another one w of the same order n over that alphabet enlarged with a new symbol, such that v is a subsequence of w and for any $n + 2k - 1$ consecutive symbols in w there is at least one occurrence of the new symbol.

For example, given this Bruijn sequence of order 3 over alphabet $\{0, 1\}$,

$$[11000101]$$

the following de Bruijn sequence of order 3 over alphabet $\{0, 1, 2\}$,

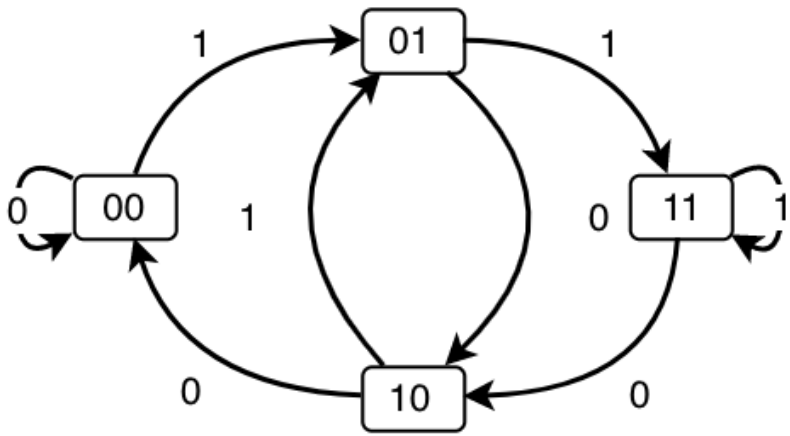
$$[1\ 2221211\ 100\ 220200\ 01\ 201021\ 01]$$

satisifes that v is a subsequence of w and given any $n + 2k - 1 = 6$ consecutive symbols in w there is at least one occurrence of the symbol 2.

de Bruijn graph

We use the terms word and sequence interchangeably.

A de Bruijn graph $G(k, n)$ is a directed graph whose vertices are the words of length n over a k-symbol alphabet and whose edges are the pairs (v, w) where $v = au$ and $w = ub$, for some word u of length $n - 1$ and possibly two different symbols a, b.



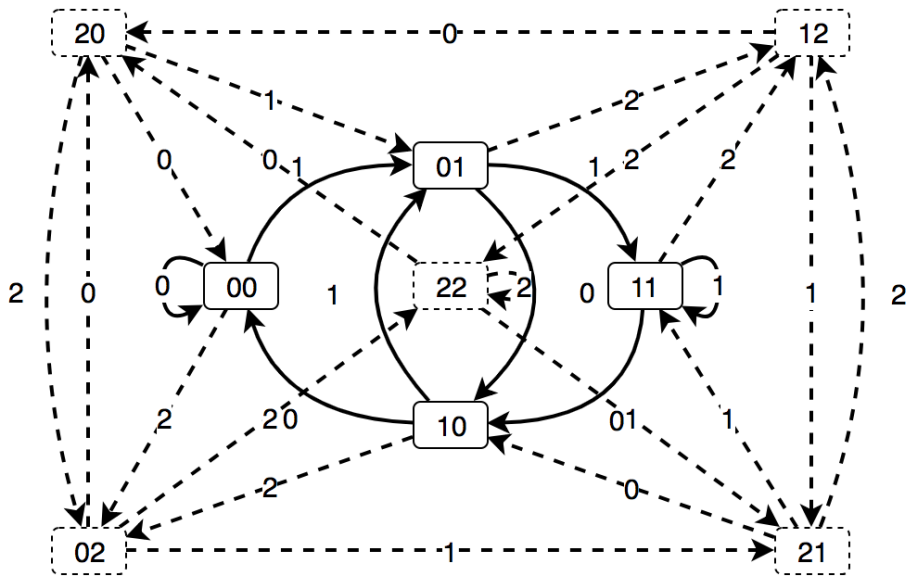
The graph $G(k, n)$ has k^n vertices and k^{n+1} edges, it is strongly connected and every vertex has the same in-degree and out-degree.

Each de Bruijn sequence of order $n + 1$ over a k-symbol alphabet can be constructed as an Eulerian cycle in $G(k, n)$.

Augmenting graph

Observe that graph $G(k, n)$ is a subgraph of $G(k+1, n)$.

The Augmenting graph $A(k+1, n)$ is the directed graph (V, E) where V is the set of length-n words over the alphabet enlarged by a new symbol s, and E is the set of pairs (v, w) such that $v = au$, $w = ub$ for some word u of length $n-1$ and symbols a, b, and either v or w have at least one occurrence of the symbol s.



Graph $G(2, 2)$ is in solid lines. The Augmenting graph $A(3, 2)$ consists of all the vertices and just the dashed lines.

Proof strategy

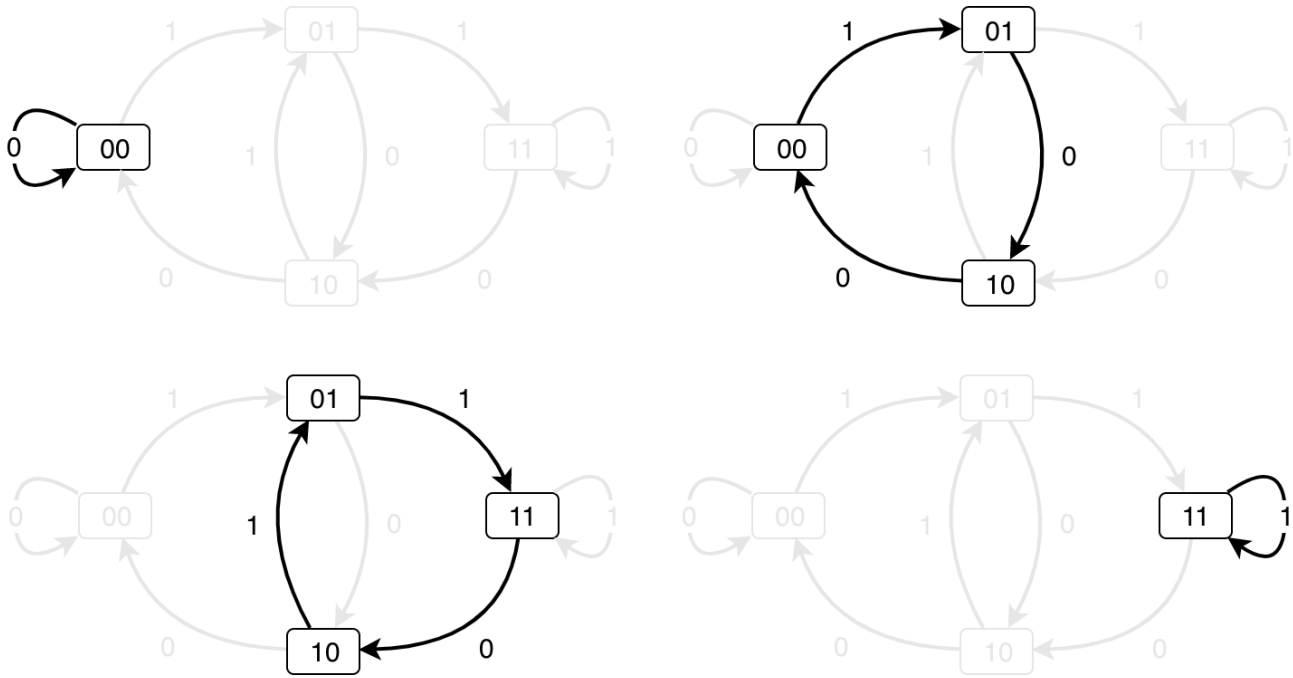
In $A(k+1,n)$ each of the vertices in $G(k,n)$ has exactly one incoming edge and exactly one outgoing edge. This outcoming edge is always labelled with the new symbol s.

To prove the Theorem construct an Eulerian cycle in $G(k+1,n)$ by joining the given Eulerian cycle in $G(k,n)$ with disjoint cycles of the augmenting graph $A(k+1,n)$ that we call petals.

Partition of $G(k, n)$ in disjoint cycles

Proposition

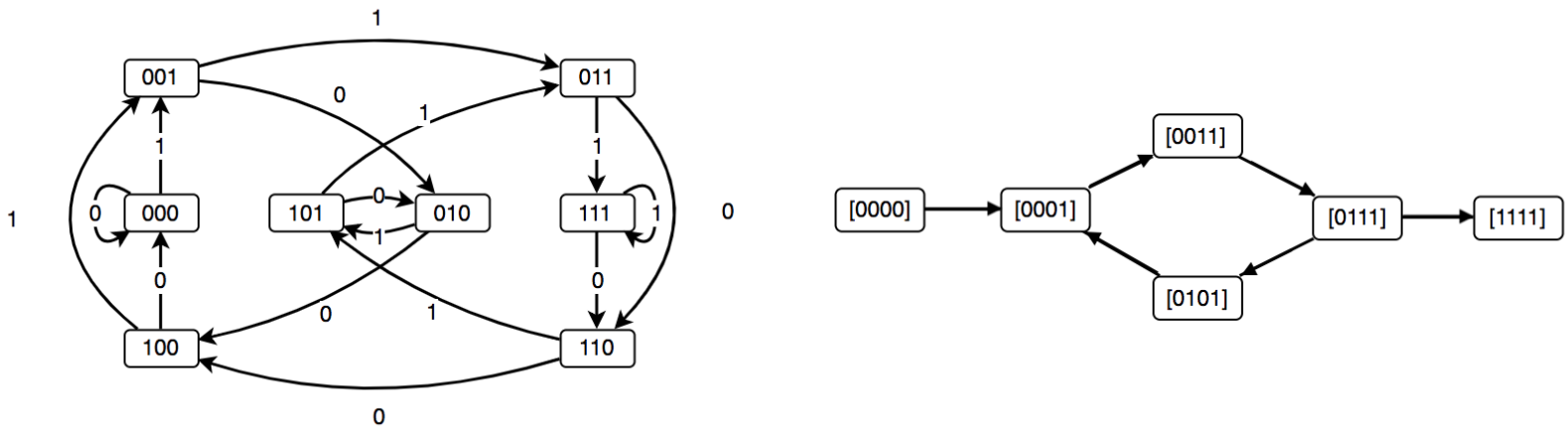
For every, k and n, the set of edges in $G(k, n)$ can be partitioned into a disjoint set of cycles identified by the circular words of length $n + 1$.



For alphabet $\{0, 1\}$ there are 4 circular words of length 3:
$[000]$, $[100]$, $[110]$ and $[111]$, each corresponds to a simple cycle in $G(2, 2)$

Definition (Graph of circular words)

For every k and n, $C(k, n+1)$ is the graph whose vertices are the circular words of length $n+1$ over the k-symbol alphabet and two vertices $[v]$ and $[w]$ are connected if there is a word u of length n and symbols a, b such that $[au] = [v]$, $[ub] = [w]$.

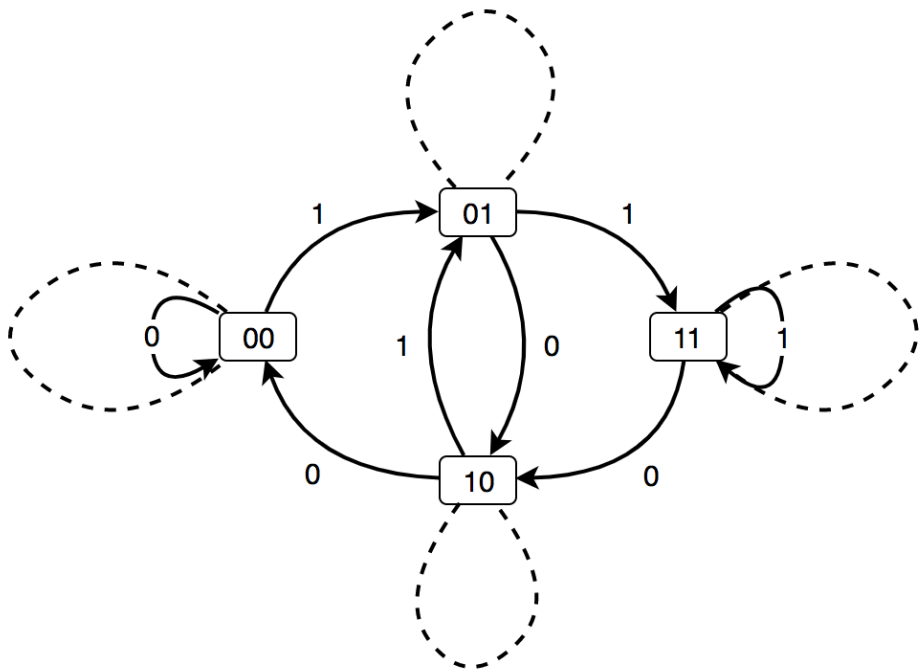


On the left $G(2, 3)$. On the right $C(2, 4)$

Petals

Definition (Petal for a vertex in $G(k, n)$)

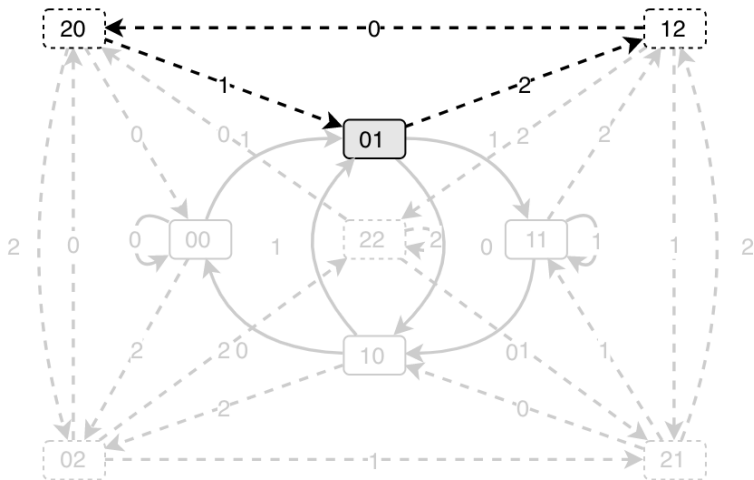
Let $\widetilde{C}(k+1, n+1)$ be the subgraph of $C(k+1, n+1)$ whose set of vertices are the circular words of length $n+1$ with at least one occurrence of symbol s. A petal for a vertex v in $G(k, n)$ is a subgraph of $\widetilde{C}(k+1, n+1)$ that seen as a cycle in $A(k+1, n)$, traverses exactly one vertex in $G(k, n)$, the vertex v.

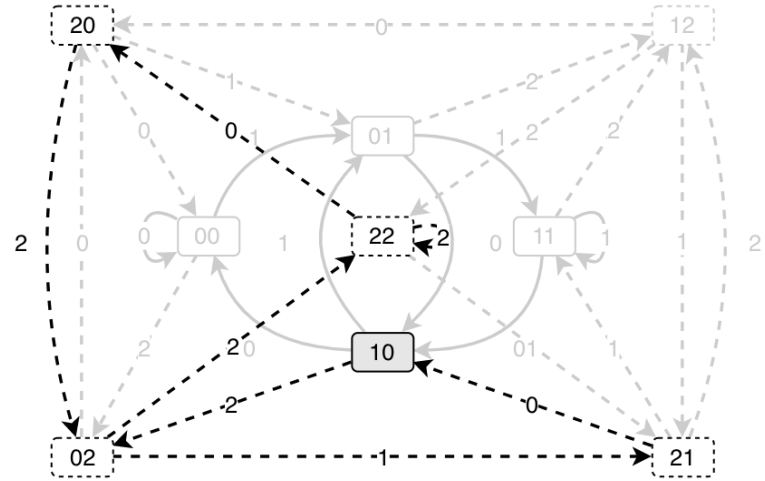


Petals for the vertices in $G(2, 2)$.

Petals are paths in the graph of circular words

There is exactly one petal for each vertex v in $G(k, n)$ and this petal starts at the circular word $[vs]$, where s is the new symbol.



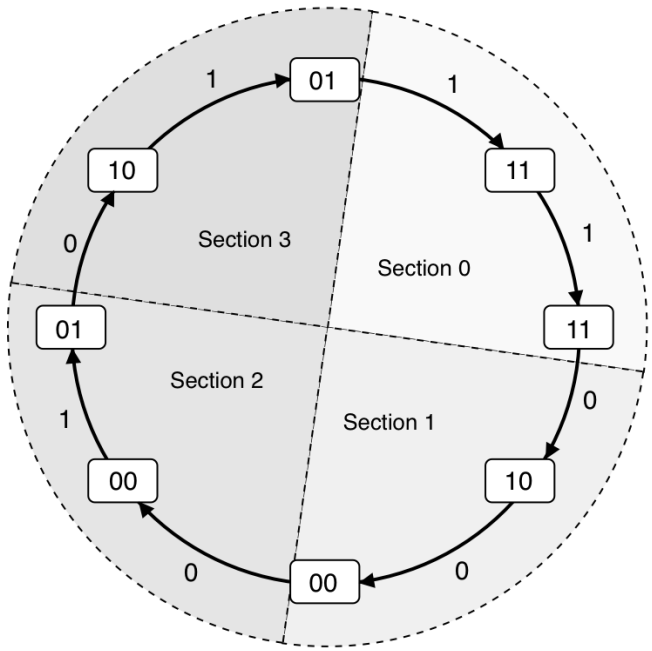


The petal for the vertex 01, which is just $[012]$.
The petal for the vertex 10 which consists of the path $[102]$ $[022]$, $[222]$

Fair distribution of the new symbol

Definition (Section of a cycle)

For a pointed Eulerian cycle in $G(k,n)$ given by the sequence of edges $e_1, \ldots e_{k^{n+1}}$ and a non-negative integer j such that $0 \leq j < k^n$, the sequence of vertices $v_{jk}, ..., v_{jk+k-1}$, where each v_i is the head of e_i, is a section j of the cycle.

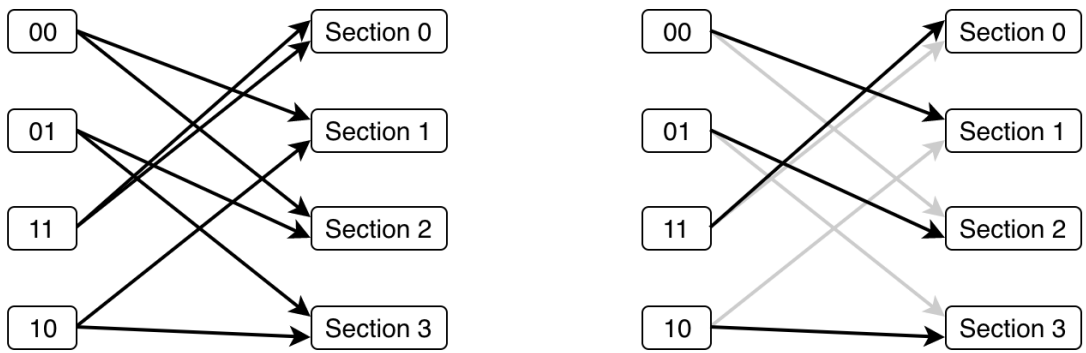


Eulerian cycle in $G(2,2)$ given by $[11000101]$ started at vertex 11 has section 0=$(11,11)$,section 1=$(10,00)$,section 2=$(00,01)$,section 3=$(10,01)$

A matching problem

The de Bruijn graph $G(k, n)$ has k^n vertices and k^{n+1} edges.
An Eulerian cycle in $G(k, n)$ has k^n sections with k vertices each section.

Since there are the same number of vertices as sections we would like to choose one vertex from each section to place a petal. The problem is that each vertex occurs k times in the Eulerian cycle but not necessarily at k different sections. We pose it as a matching problem.



Matching problem

Definition (Distribution graph)

Given pointed Eulerian cycle in $G(k,n)$ the Distribution graph $D(k,n)$ is a k-regular bipartite graph where the two vertex classes are the vertices in $G(k,n)$ and the sections of the Eulerian cycle and there is an edge (v,j) if v belongs to the section j.

A matching in a graph D is a set of edges such that no two edges share a common vertex. A vertex is matched if it is an endpoint of one of the edges in the matching. A perfect matching is a matching that matches all vertices in the graph.

Hall's marriage theorem

Lemma

For every Distribution graph $D(k, n)$ there is a perfect matching.

Let D be a finite bipartite graph consisting of two disjoint sets of vertices X and Y with edges that connect a vertex in X to a vertex in Y.

For a subset W of X, let $N(W)$ be the set of all vertices in Y adjacent to some element in W . Hall's marriage theorem (1935) states that there is a matching that entirely covers X if and only if for every subset W in$X$, $|W| \leq |N(W)|$.

Consider a Distribution graph $D(k, n)$ and call X to the set of vertices $G(k, n)$ and Y to the set of sections. For any $W \subseteq X$ such that $|W| = r$, the sum of the out-degree of these r vertices is rk. Given that the in-degree for any vertex in Y is k, we have that $|N(W)| \geq r$. Then, there is a matching that entirely covers X.

Furthermore, since the number of vertices is equal to the number of sections, $|X| = |Y|$ and the matching is perfect.

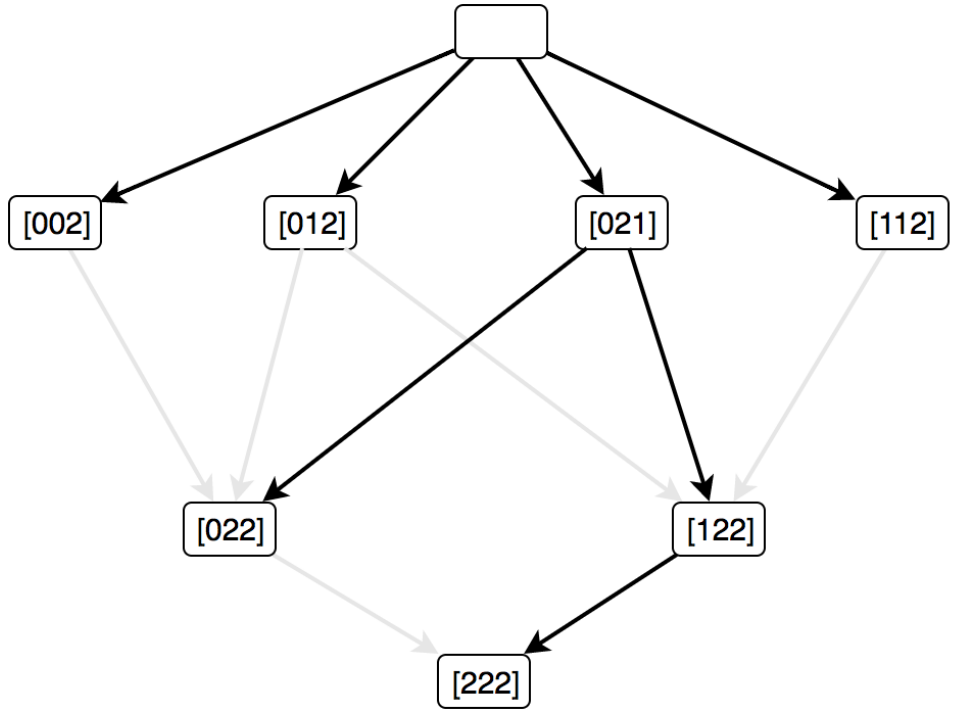
Perfect matching

To obtain a perfect matching in a Distribution graph we can use any method to compute the maximum flow in a network, such as Edmonds-Karp algorithm.

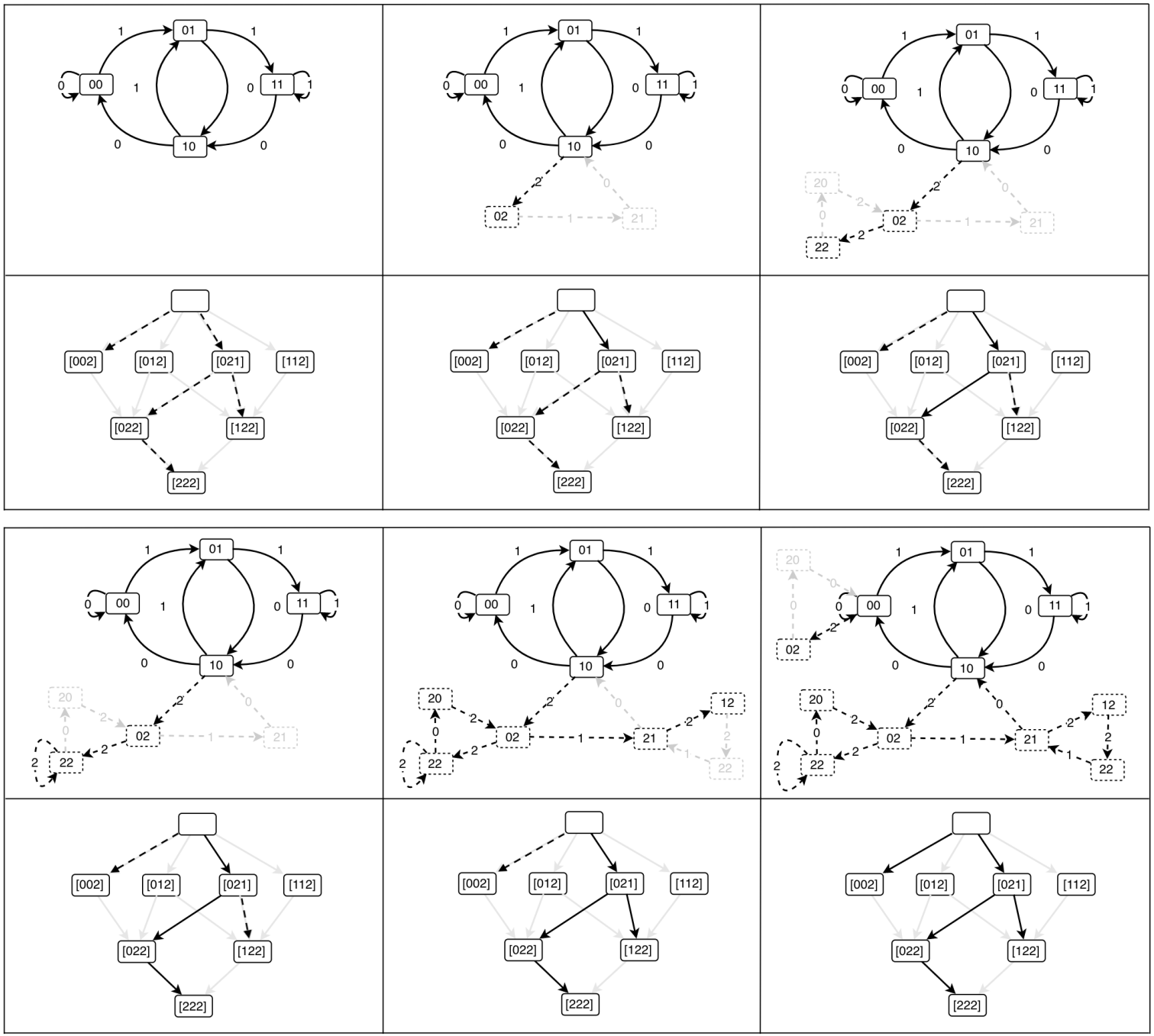
We define the flow network by adding adding two vertices to the Distribution graph, the source and the sink. Add an edge from the source to each vertex in X and add an edge from each vertex in Y to the sink.
Assign capacity 1 to each of the edges of the flow network.
The maximum flow of the network is $|X|$.
This flow has the edges of a perfect match.

Partition of the augmenting graph

We must partition the set of edges in $A(k+1, n)$ into petals.
We define a Petals tree as a root that branches out in a subgraph of $\widetilde{C}(k+1, n+1)$. It has height $n+1$, the vertices at distance d to the root have exactly d occurrences of the new symbol s, for $d = 1, \ldots, n+1$.



A petals tree with four petals, one for each vertex of $G(2, 2)$.
The search of the maximum flow is the most expensive part of the construction. Edmonds-Karp algorithm has running time



Extension of the de Bruijn sequence [11000101]

General problem

Fix an alphabet A and a new symbol s not in A.

Given an Borel normal sequence x of symbols in A transform it into a normal sequence y of symbols in the extended alphabet $A \cup \{s\}$ such that x is a subsequence of y and the speed of convergence to normality of x and y coincide.