CUCKOO HASHING REVISITED

motivated by Luc Devroye, jointly worked out with Reinhard Kutzelnigg

Michael Drmota*

Institut für Diskrete Mathematik und Geometrie

TU Wien

michael.drmota@tuwien.ac.at

www.dmg.tuwien.ac.at/drmota/

* supported by the Austrian Science Foundation FWF, grant S9600.

Contents

- Cuckoo Hashing
- Cuckoo Graph
- Random Bipartite Graphs
- Asymptotic Results
- Generating Functions
- Double Saddle Points

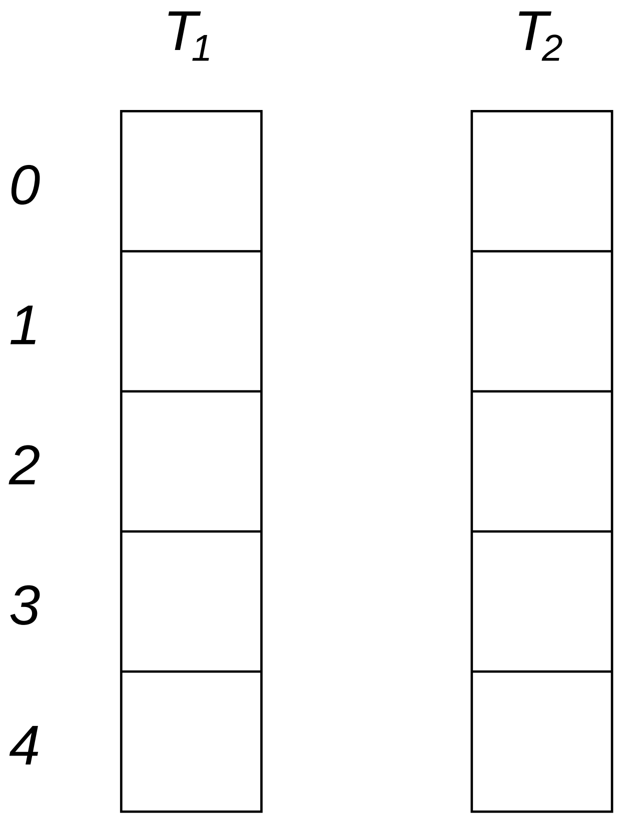
Cuckoo Hashing

[Pagh and Rodler, 2001]

- 2 tables T_1, T_2 of size m
- 2 hash functions $\boxed{h_1(x), h_2(x)}$.
- Every key x is stored at $\boxed{h_1(x) \in T_1}$ or at $\boxed{h_2(x) \in T_2}$.

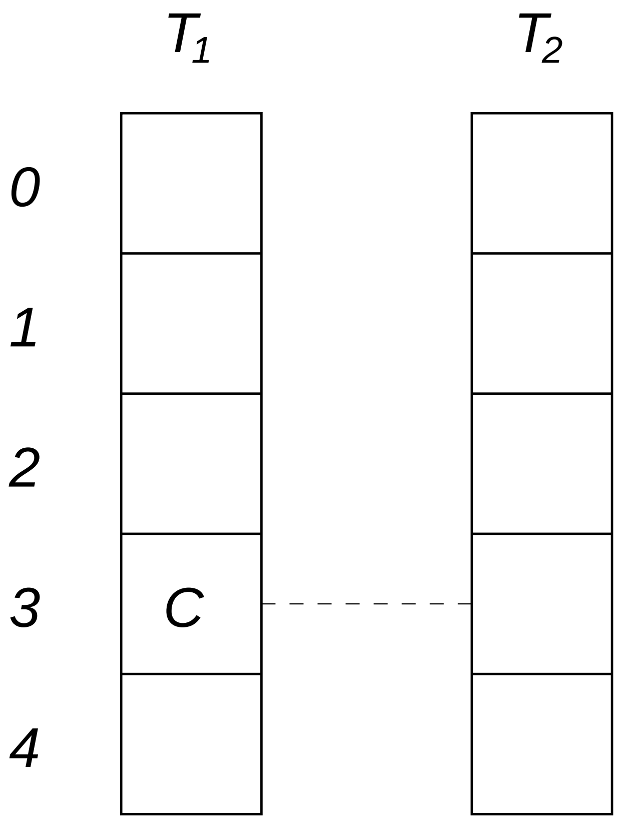
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



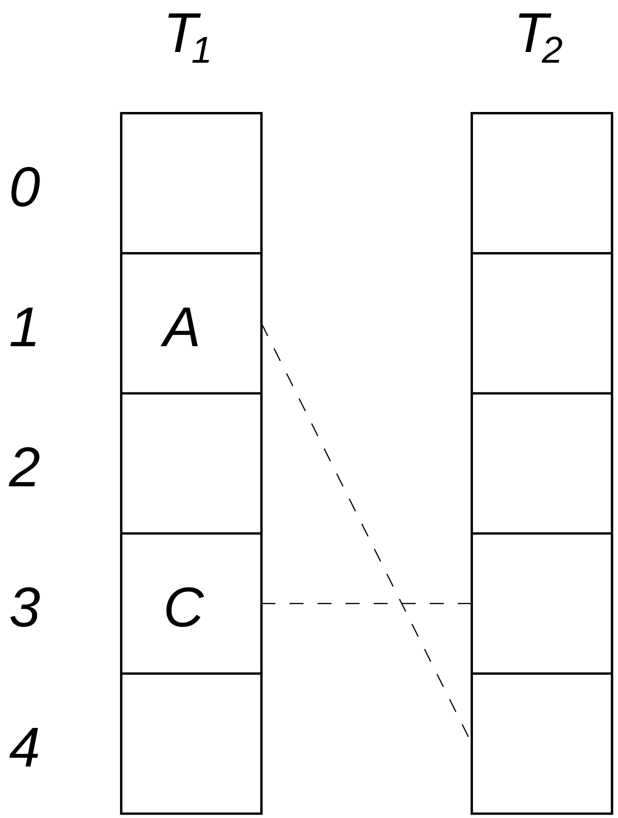
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



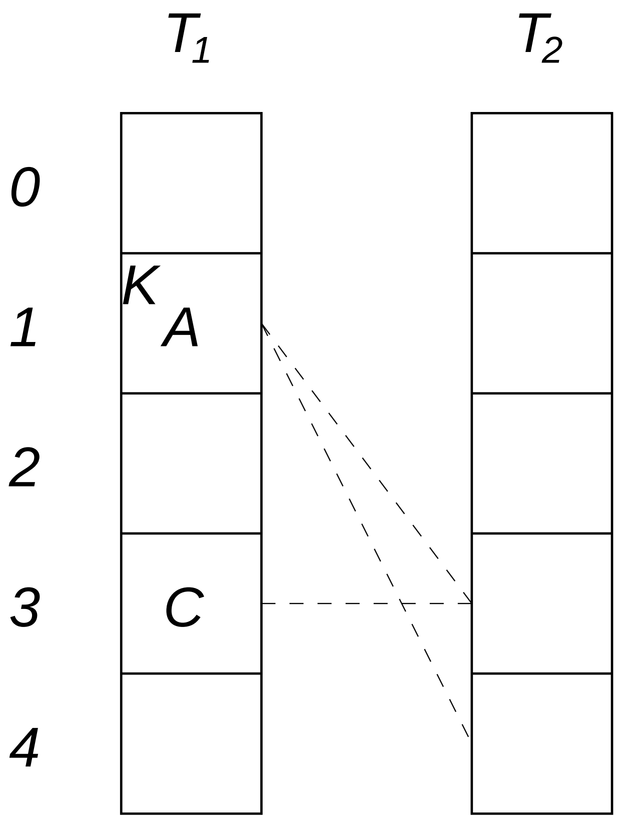
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



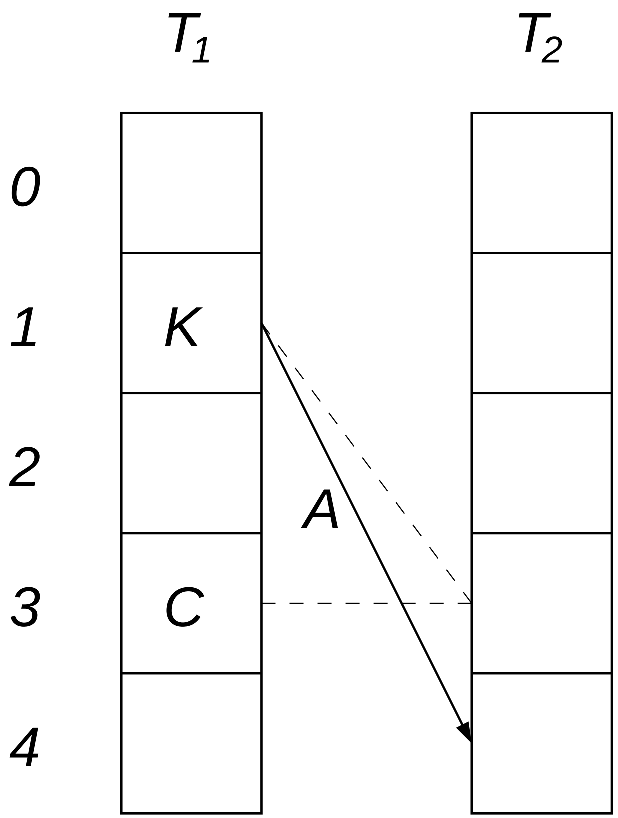
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



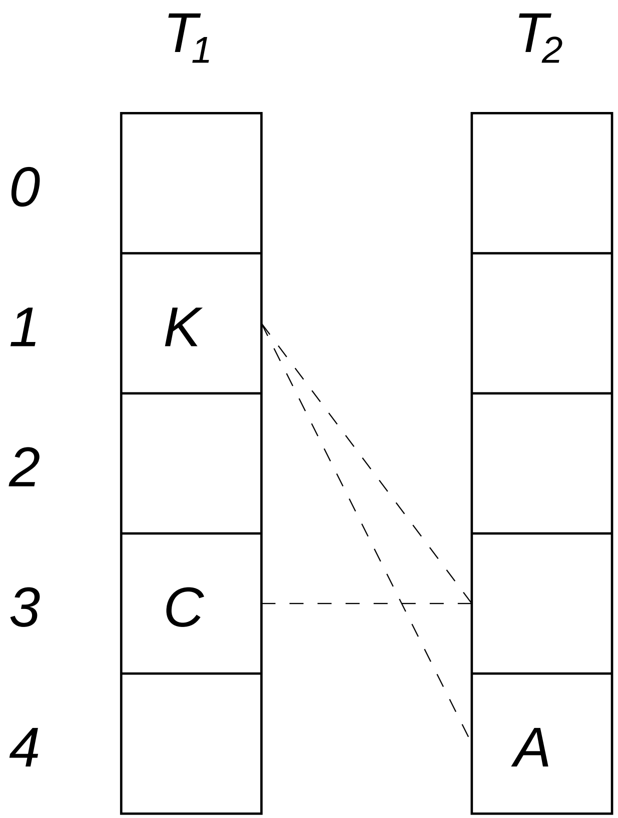
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



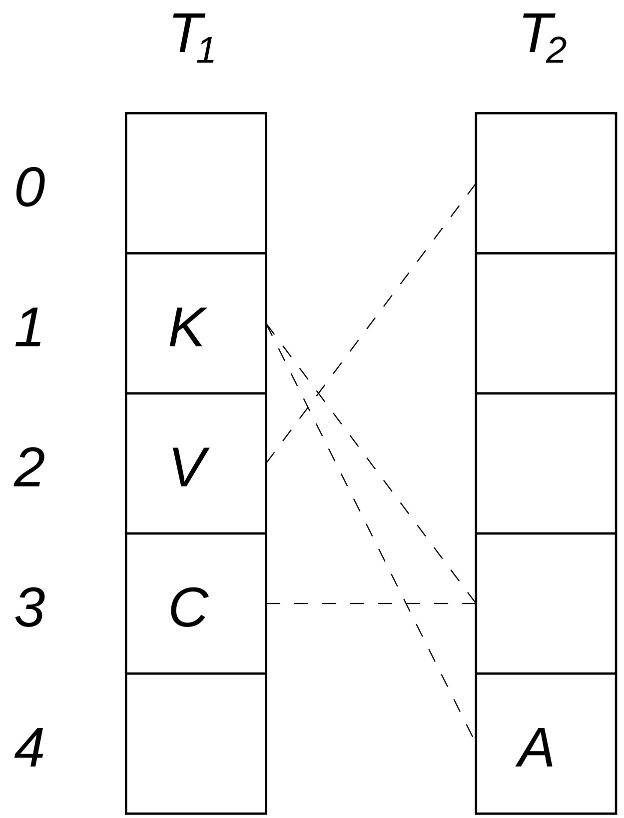
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



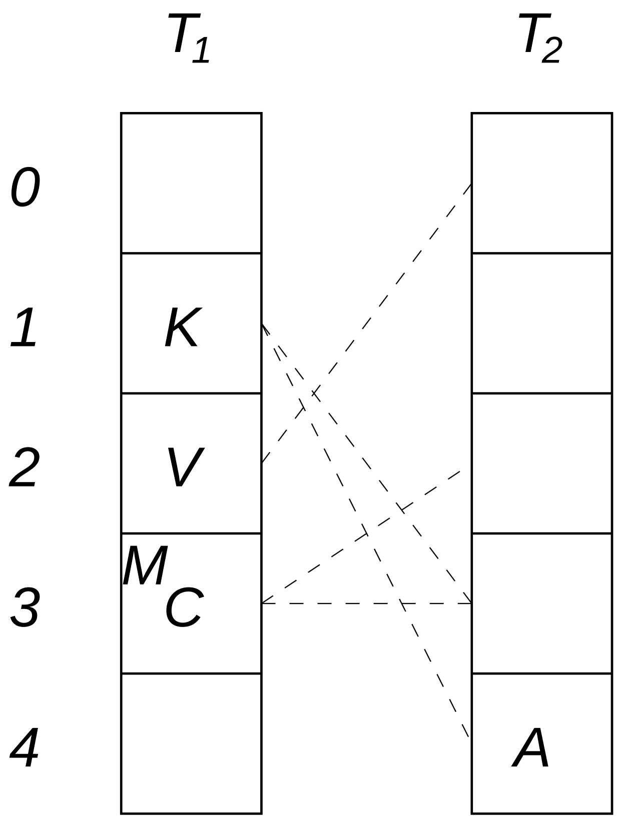
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



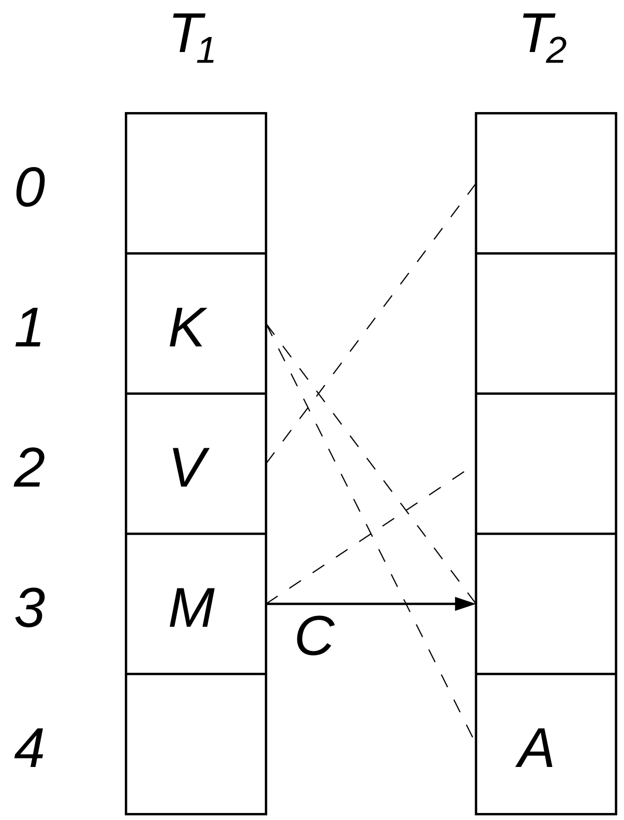
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



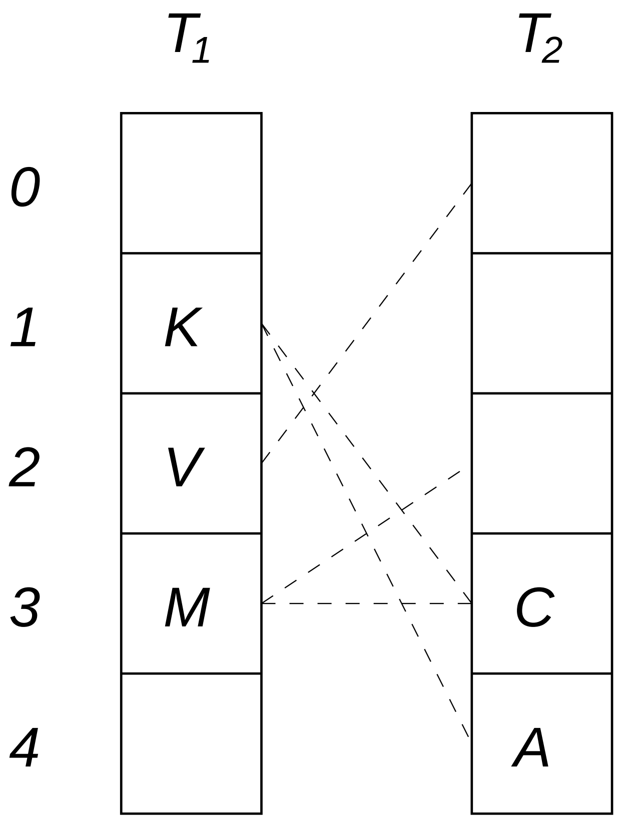
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



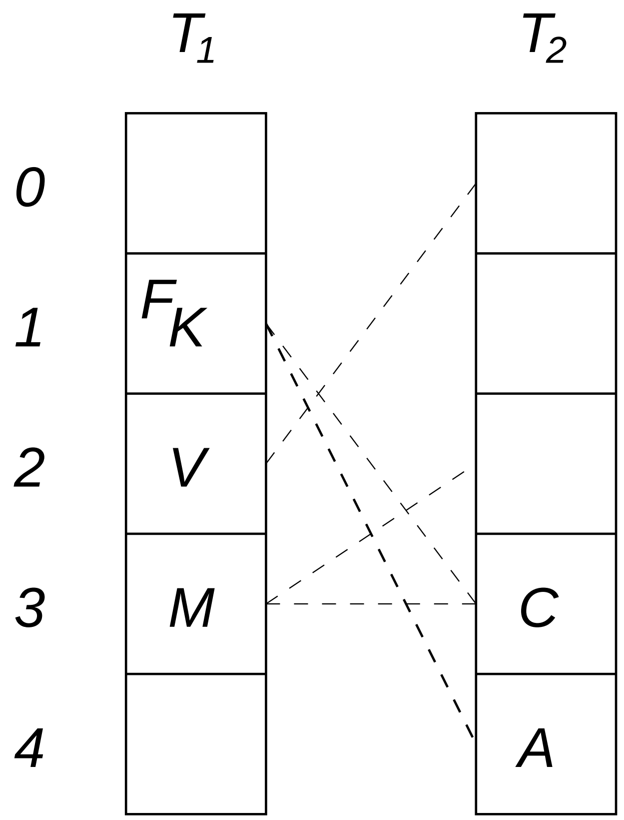
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



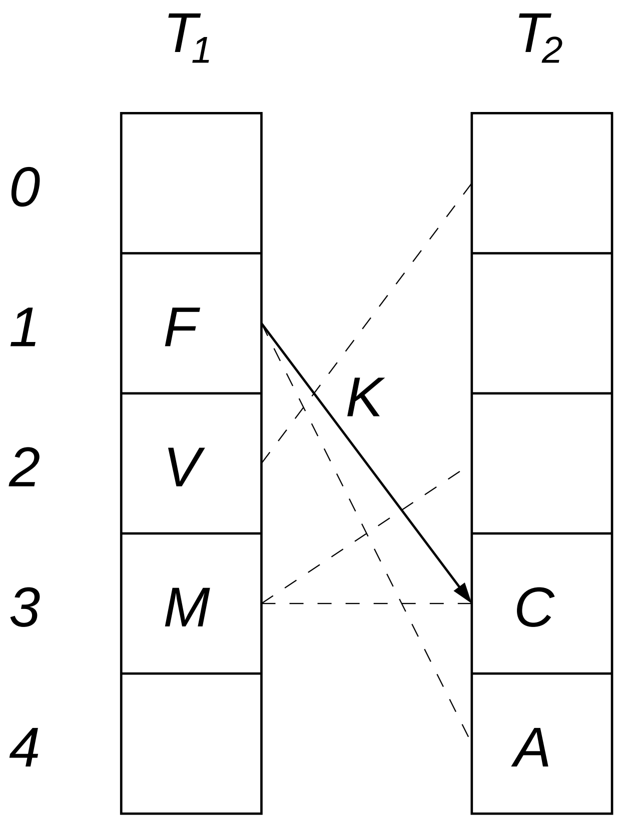
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



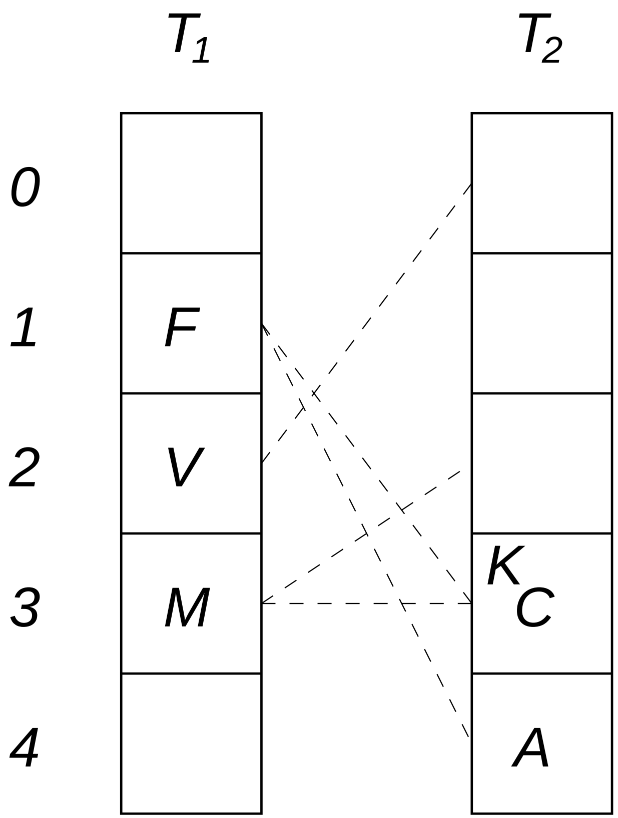
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



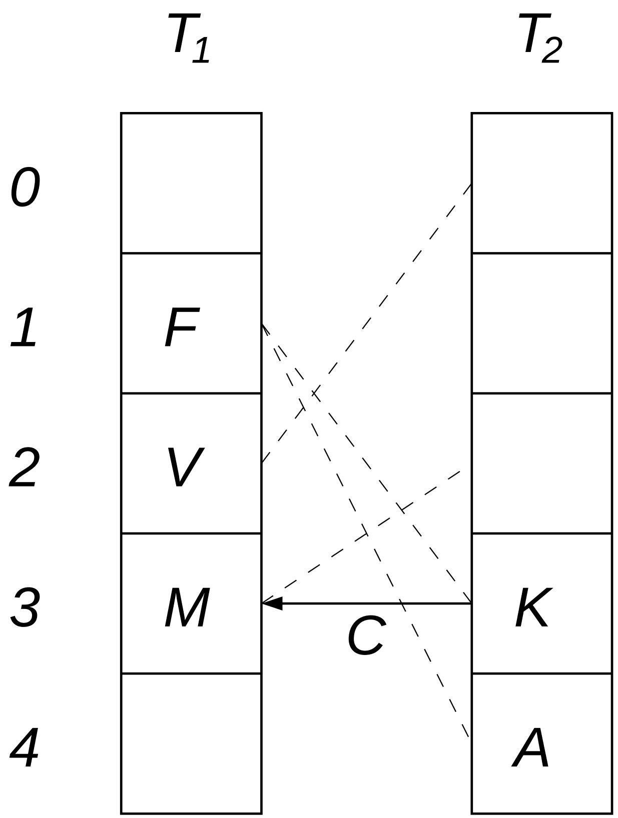
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



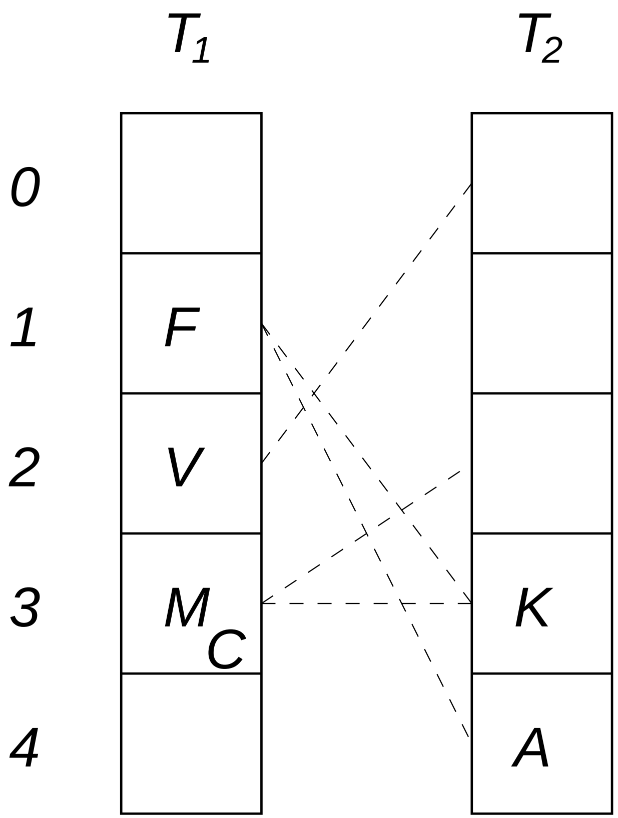
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



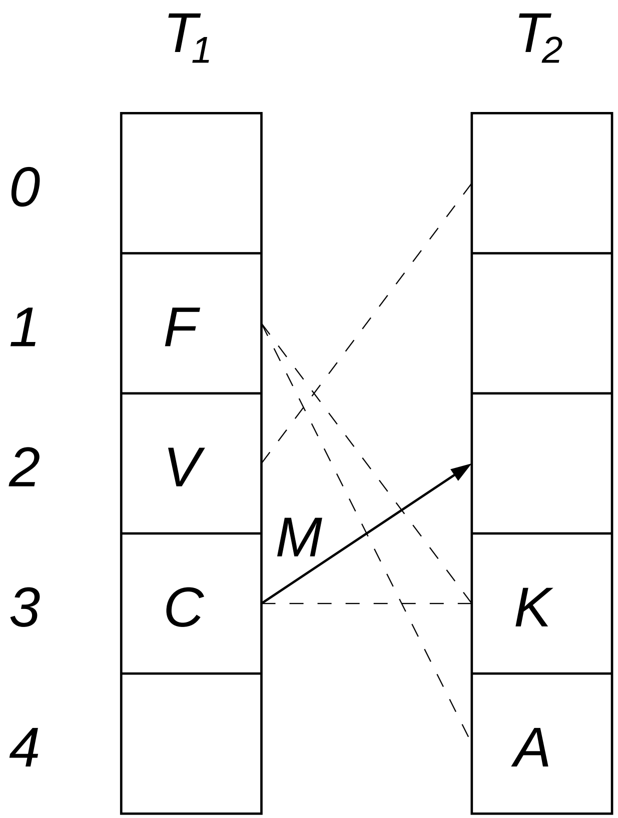
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



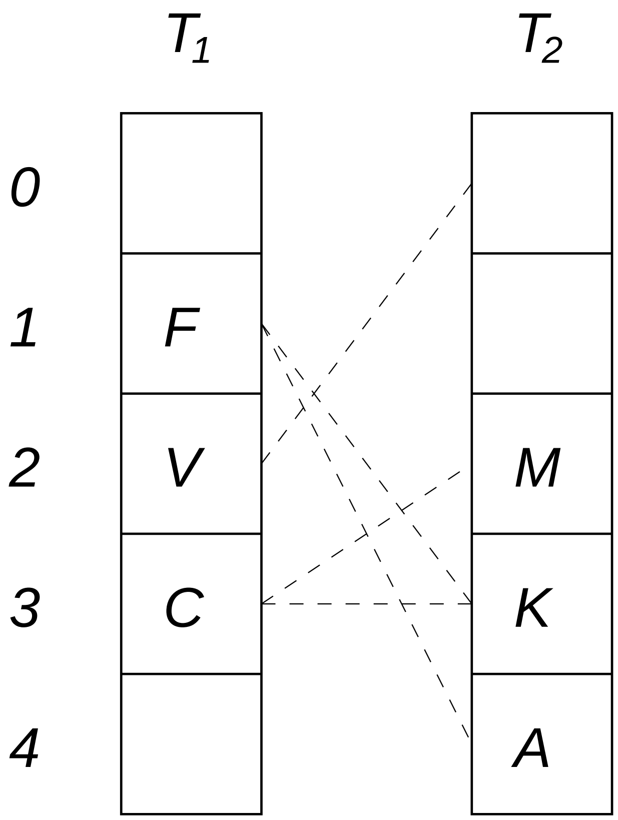
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



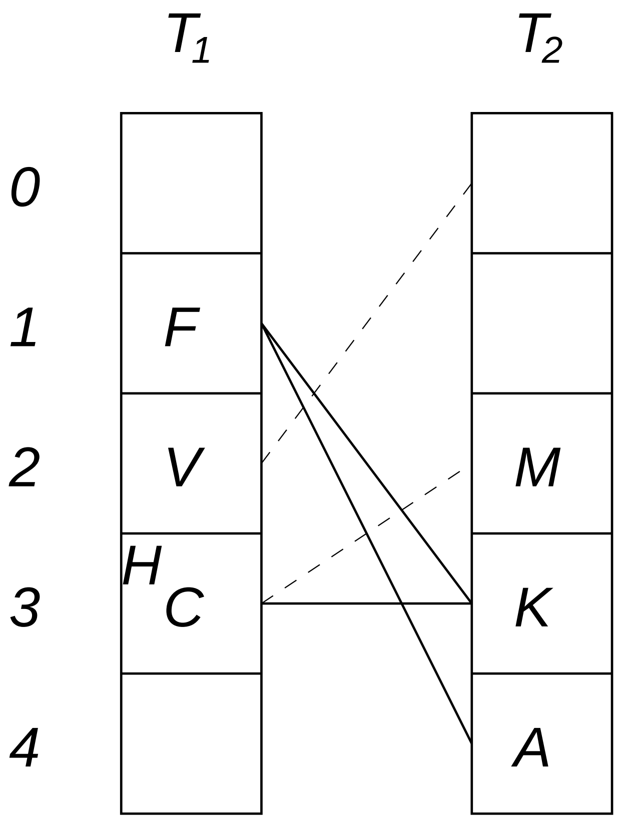
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



Cuckoo Hashing

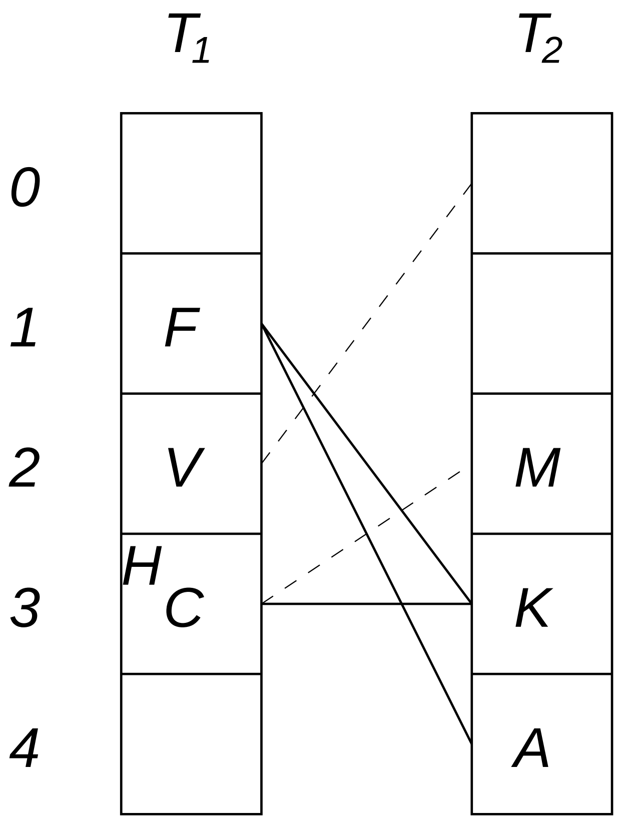
	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3



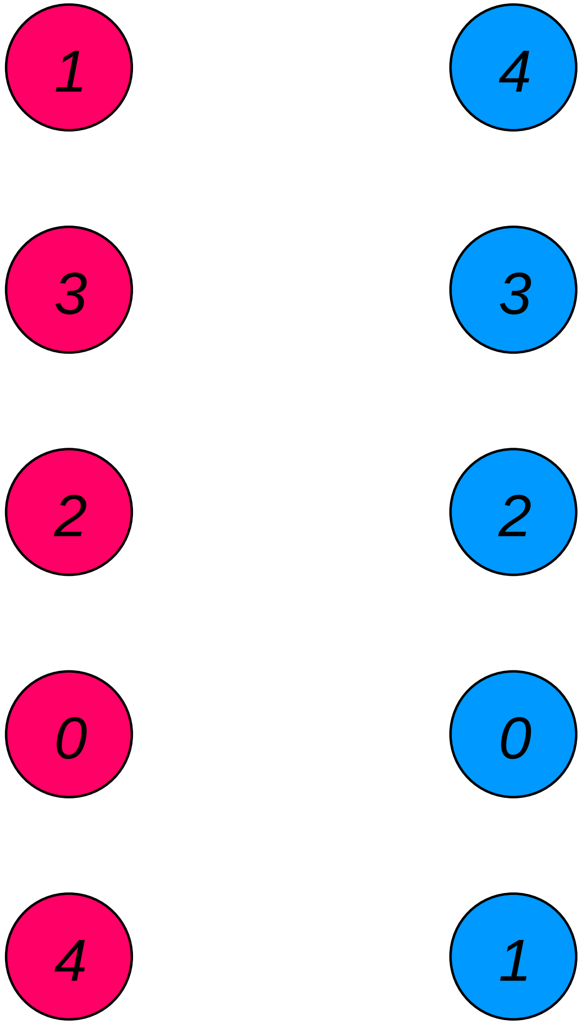
Cuckoo Hashing

	C	A	K	V	M	F	H
h_1	3	1	1	2	3	1	3
h_2	3	4	3	0	2	4	3

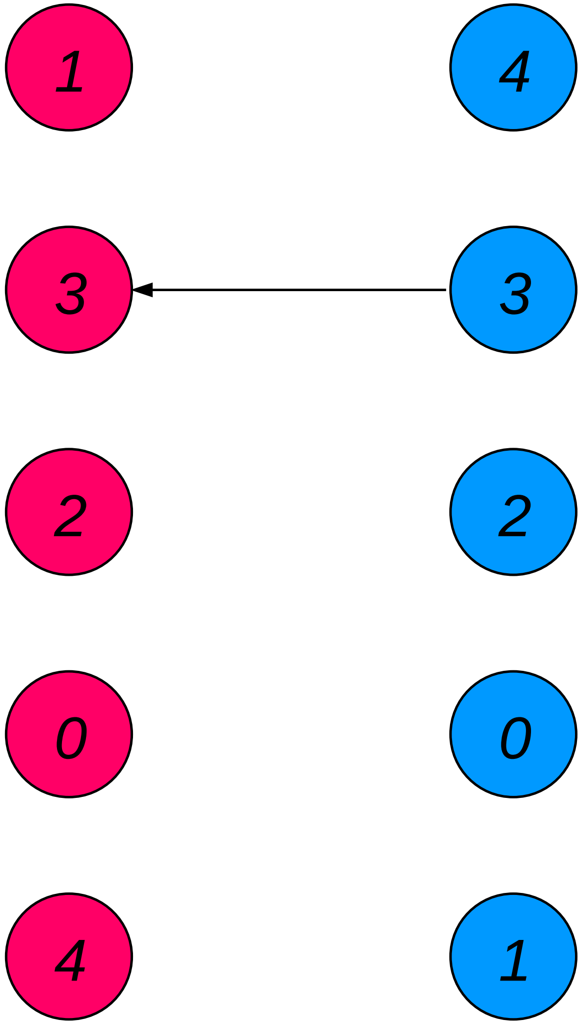
REHASH !!



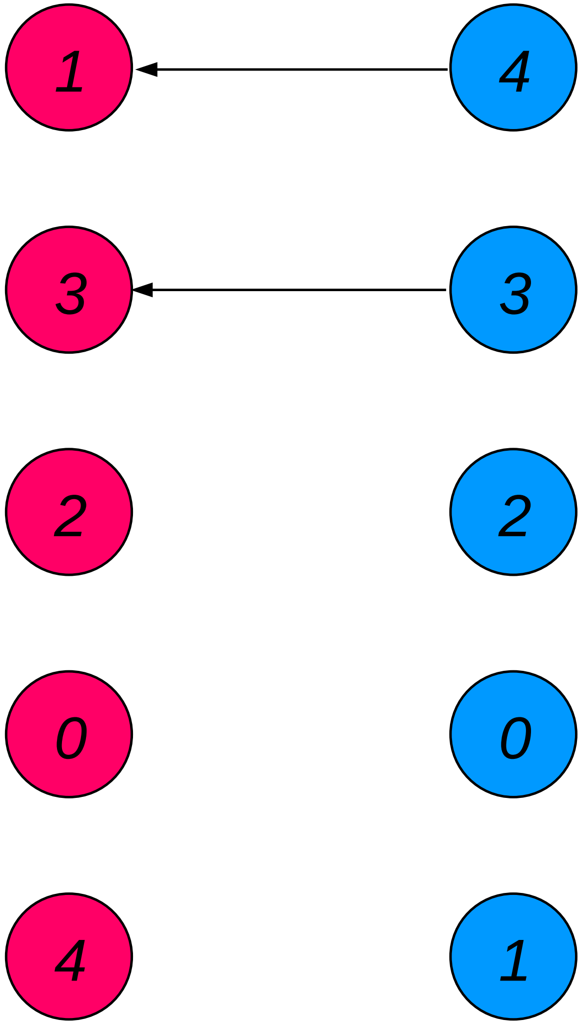
Cuckoo Graph



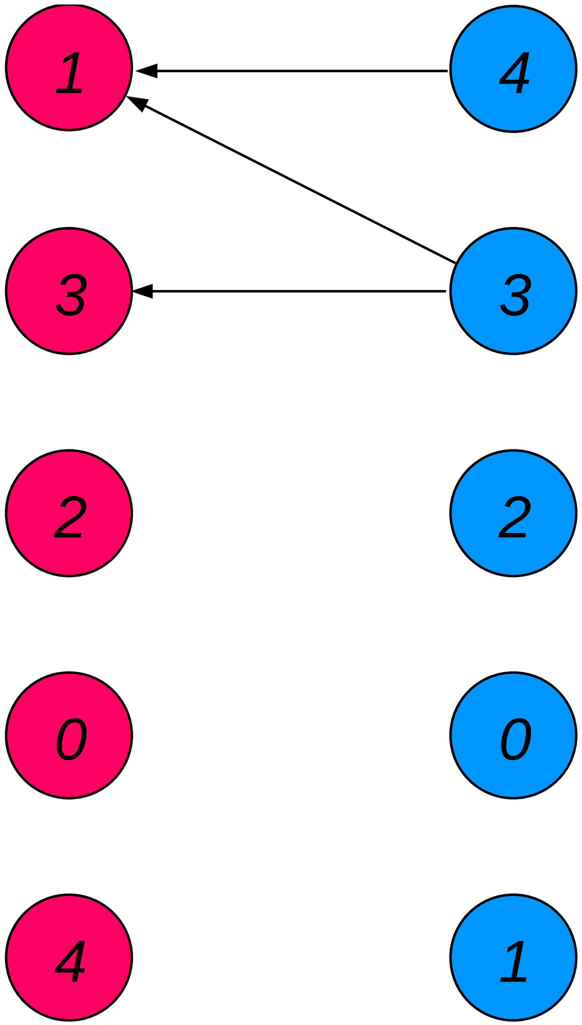
Cuckoo Graph



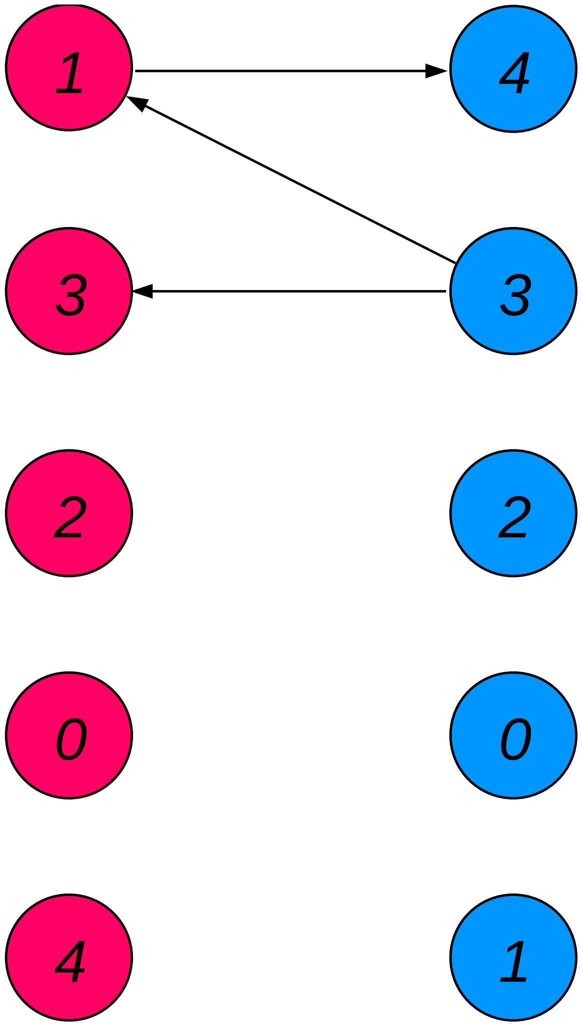
Cuckoo Graph



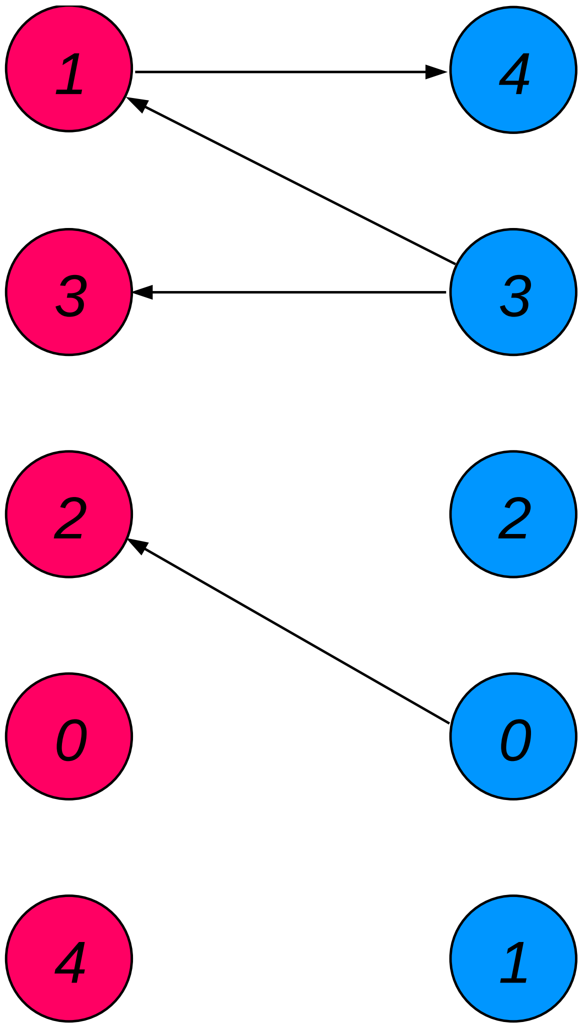
Cuckoo Graph



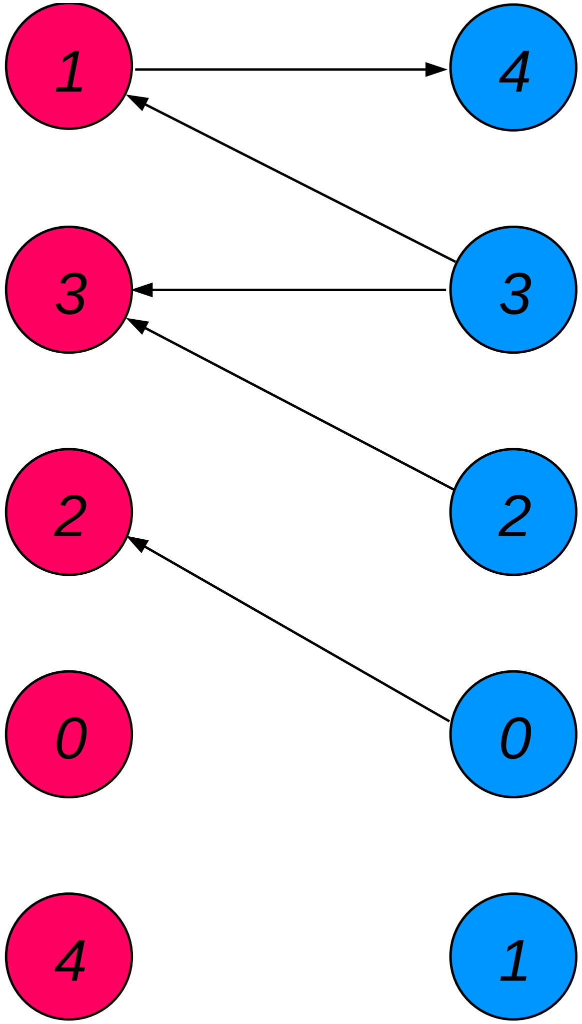
Cuckoo Graph



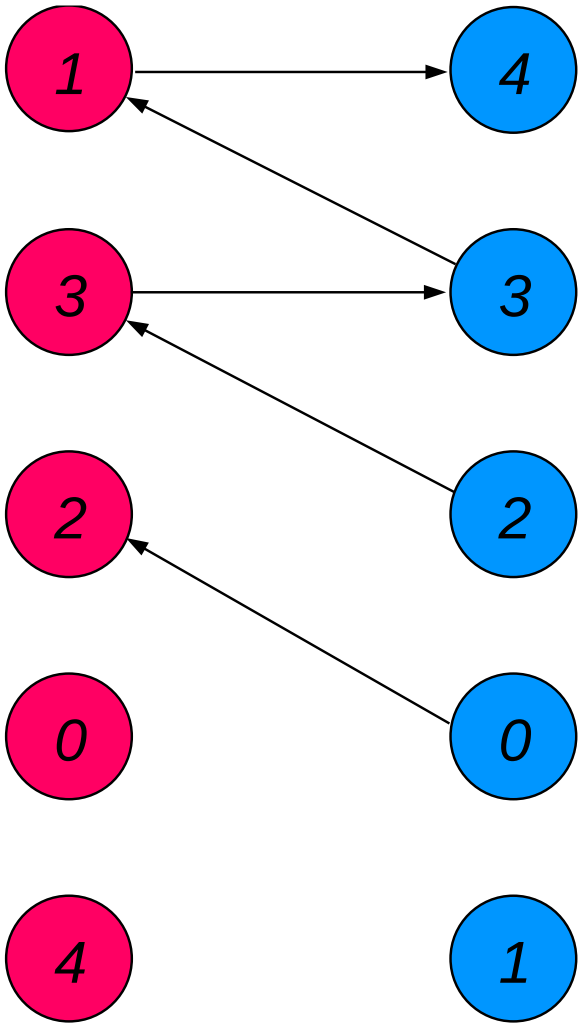
Cuckoo Graph



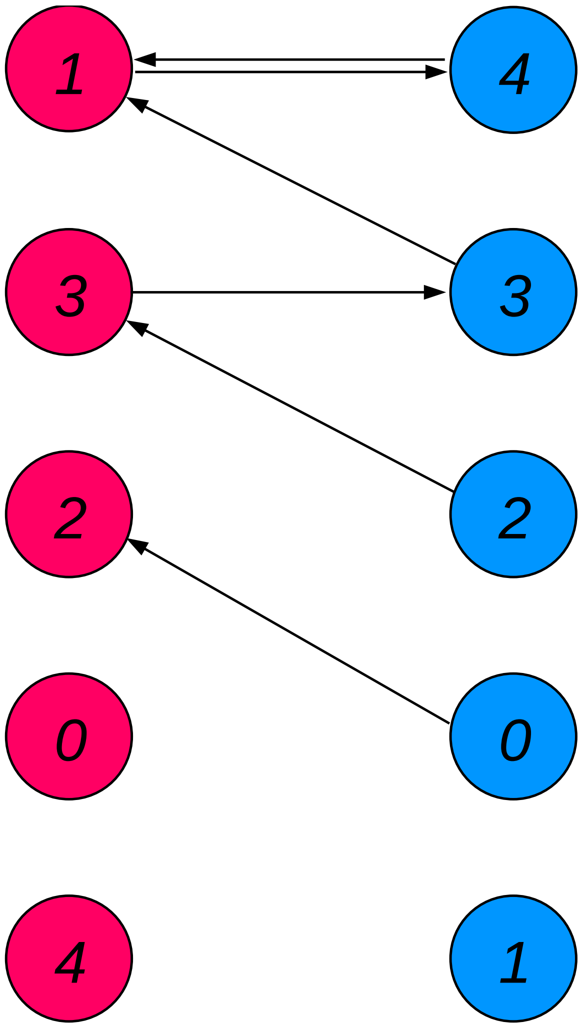
Cuckoo Graph



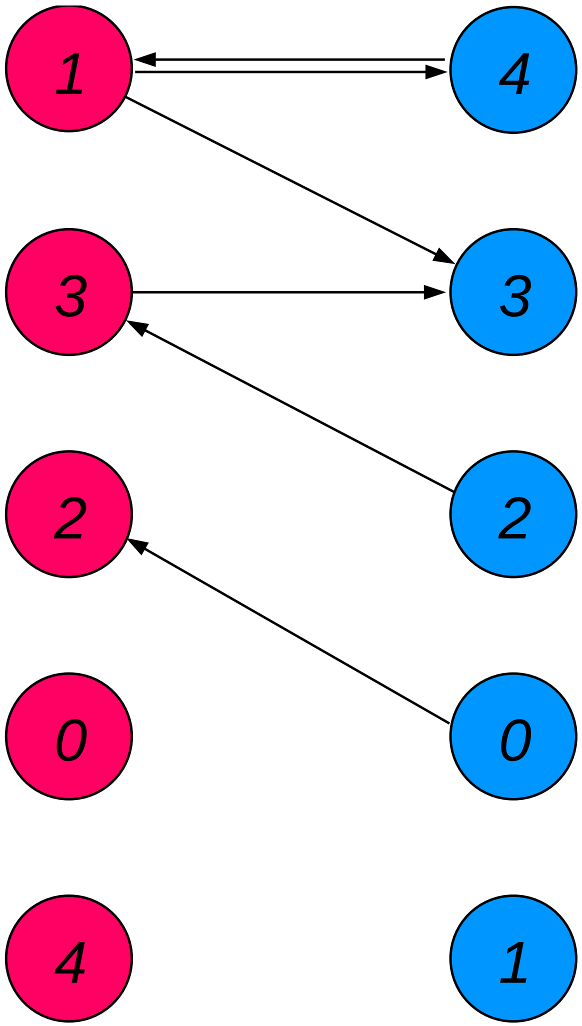
Cuckoo Graph



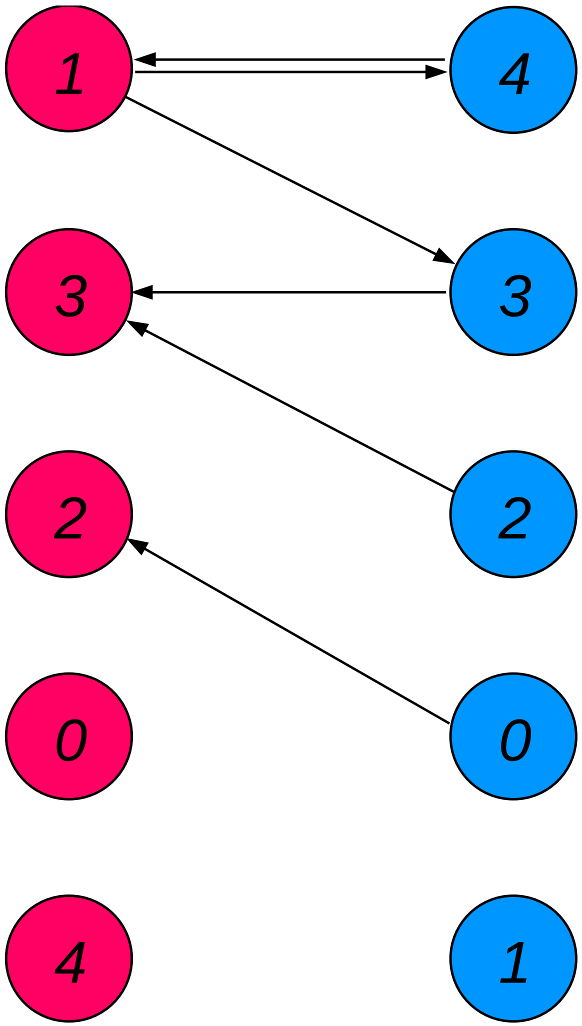
Cuckoo Graph



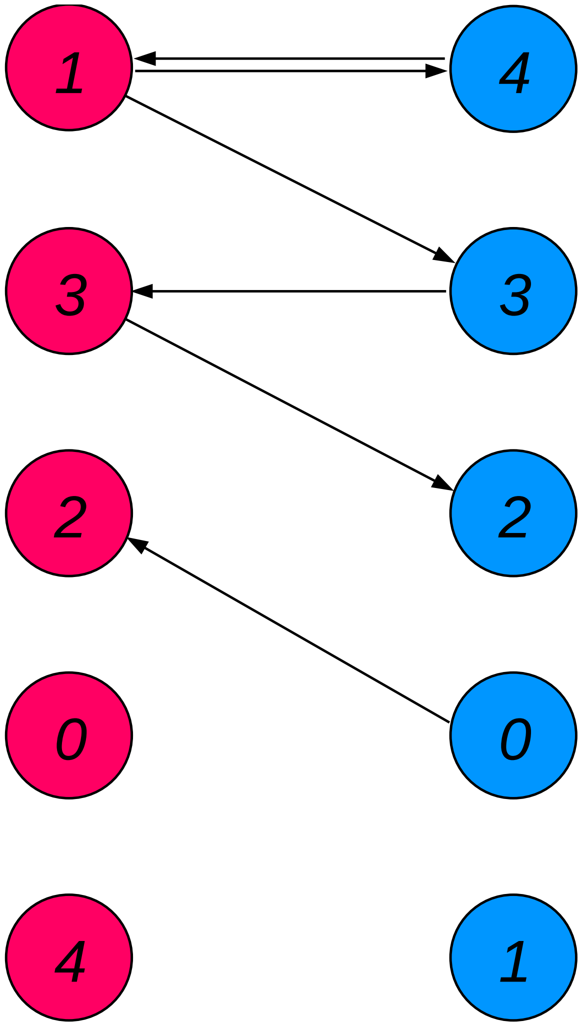
Cuckoo Graph



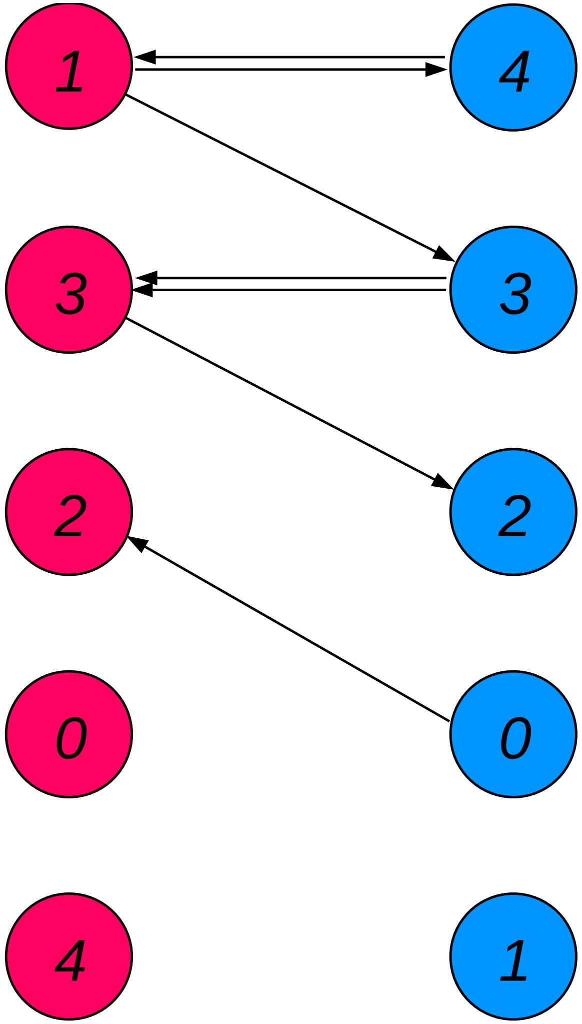
Cuckoo Graph



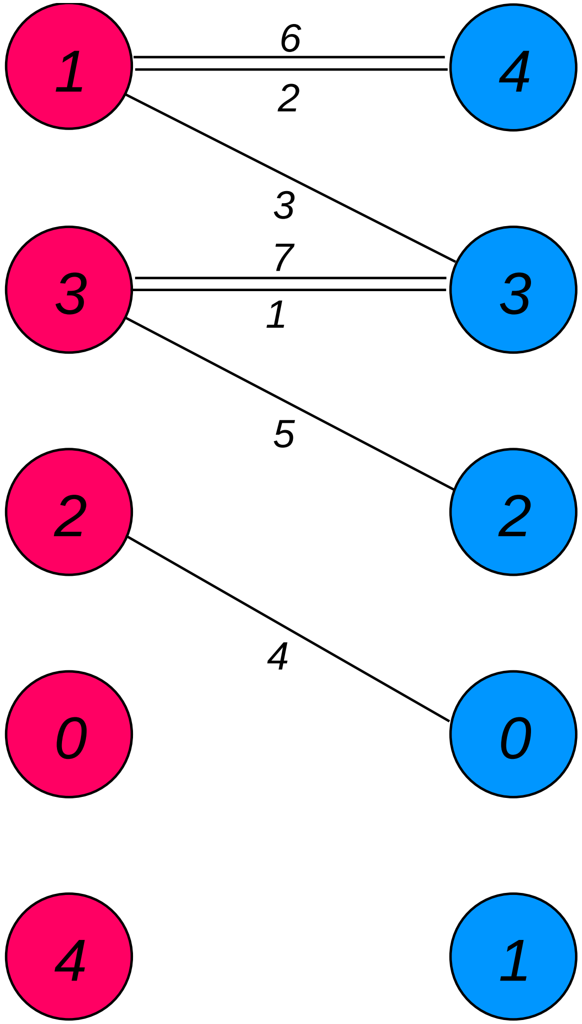
Cuckoo Graph



Cuckoo Graph



Cuckoo Graph



Cuckoo Graph

- **Bipartite Graph**

$$\boxed{G = (V_1, V_2, E)}$$

  V_1, V_2 ... *tables*, labeled vertex sets
  E ... collects hash values $e = (h_1(x), h_2(x))$, **labeled** edges

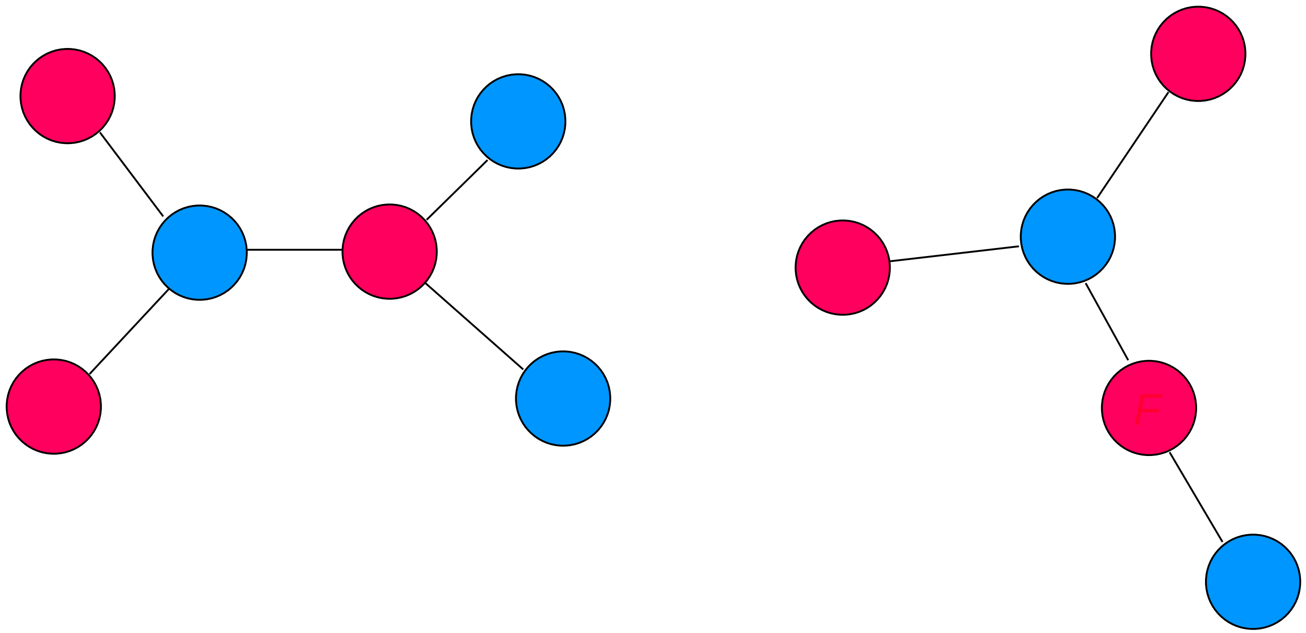
- $|V_1| = |V_2| = m$ ... table size
  $|E| = n$ ... number of keys

- Hashing works $\Longleftrightarrow$ G contains **no** complex component !!!

  (only **trees** or **unicyclic components**)

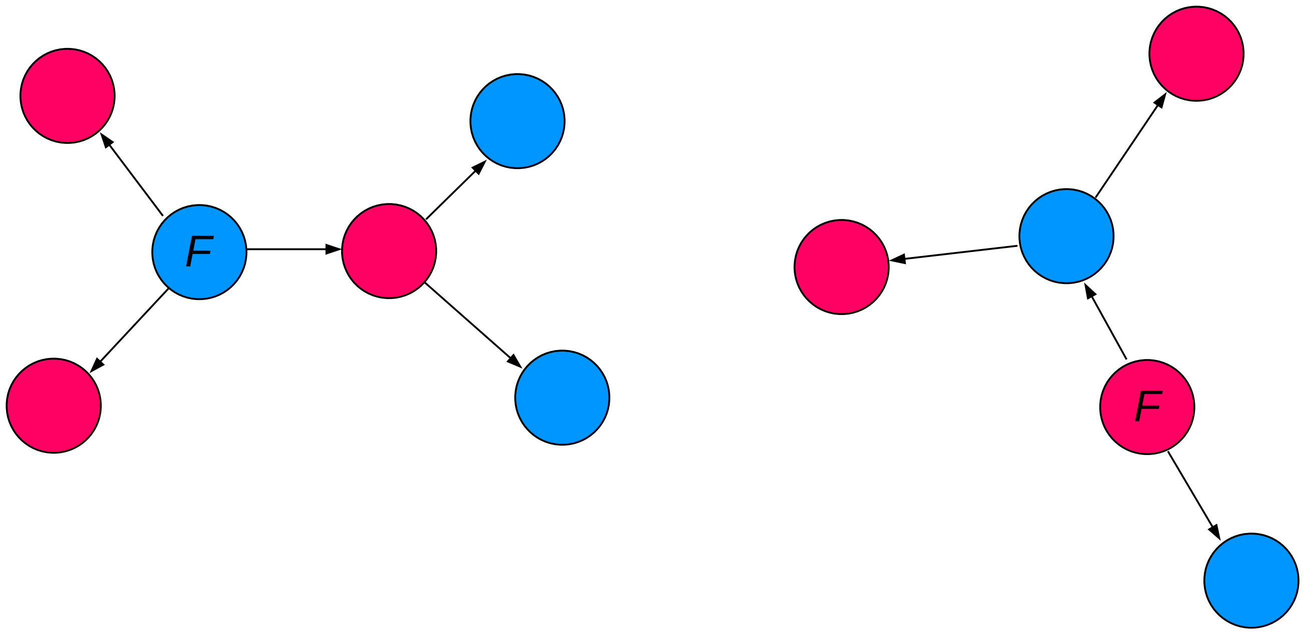
Cuckoo Graph

Edge Insertion: Good case 1: joining 2 trees



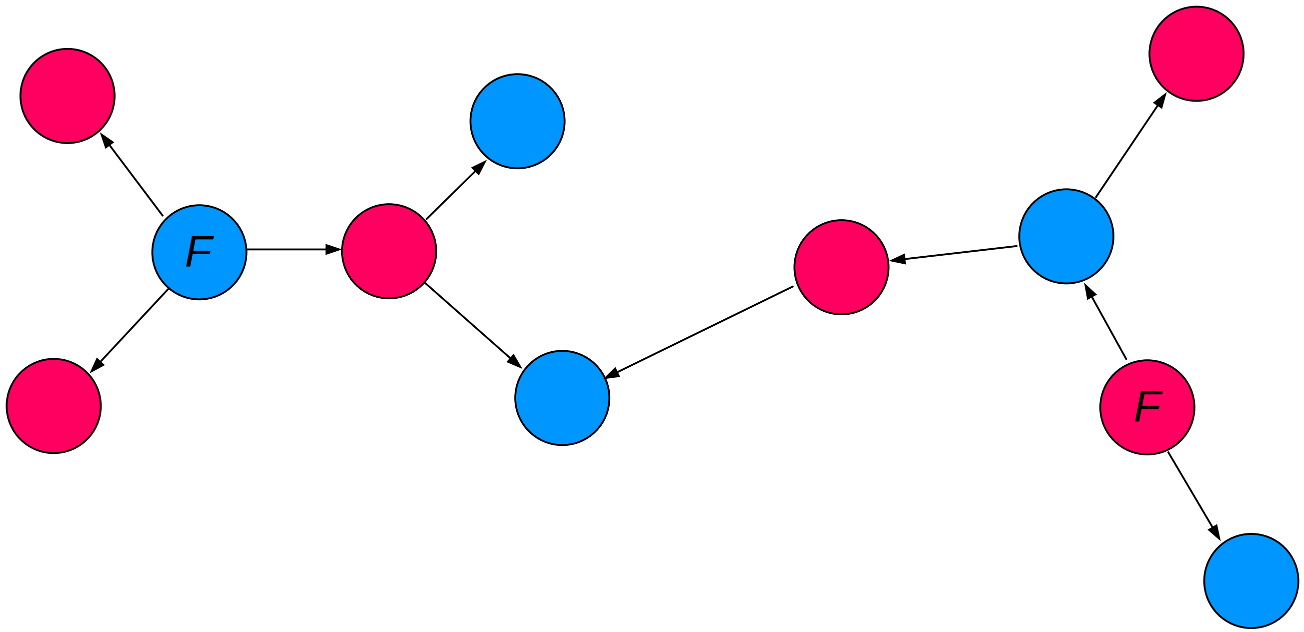
Cuckoo Graph

Edge Insertion: Good case 1: joining 2 trees



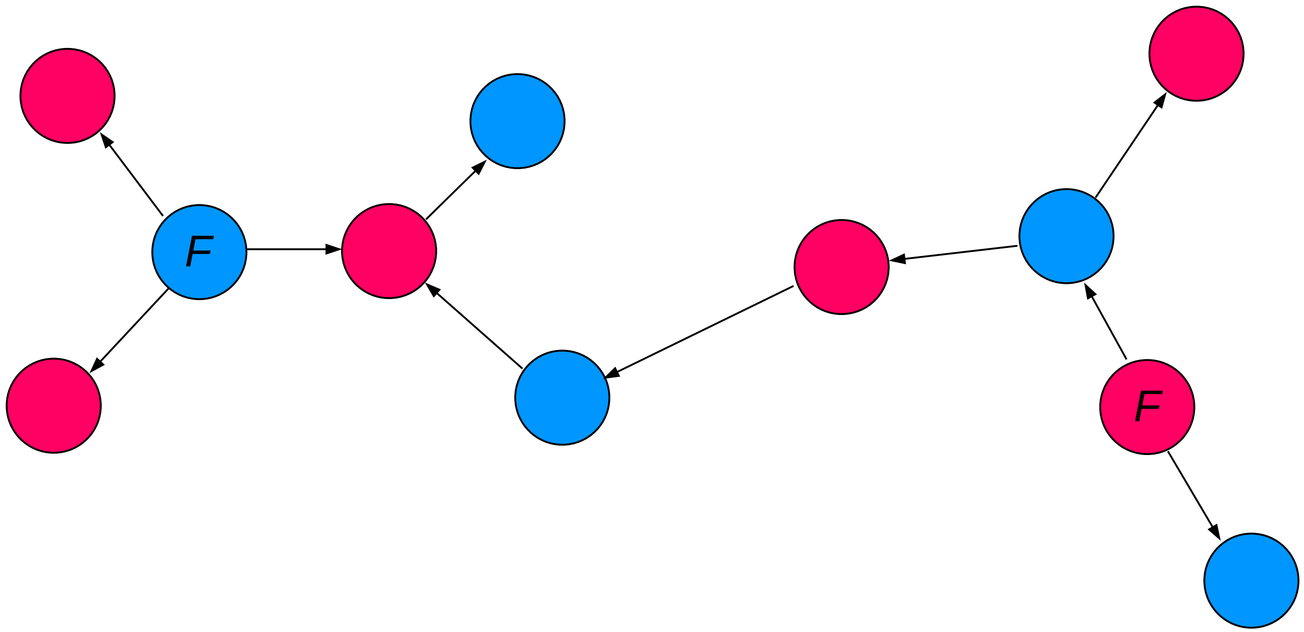
Cuckoo Graph

Edge Insertion: Good case 1: joining 2 trees



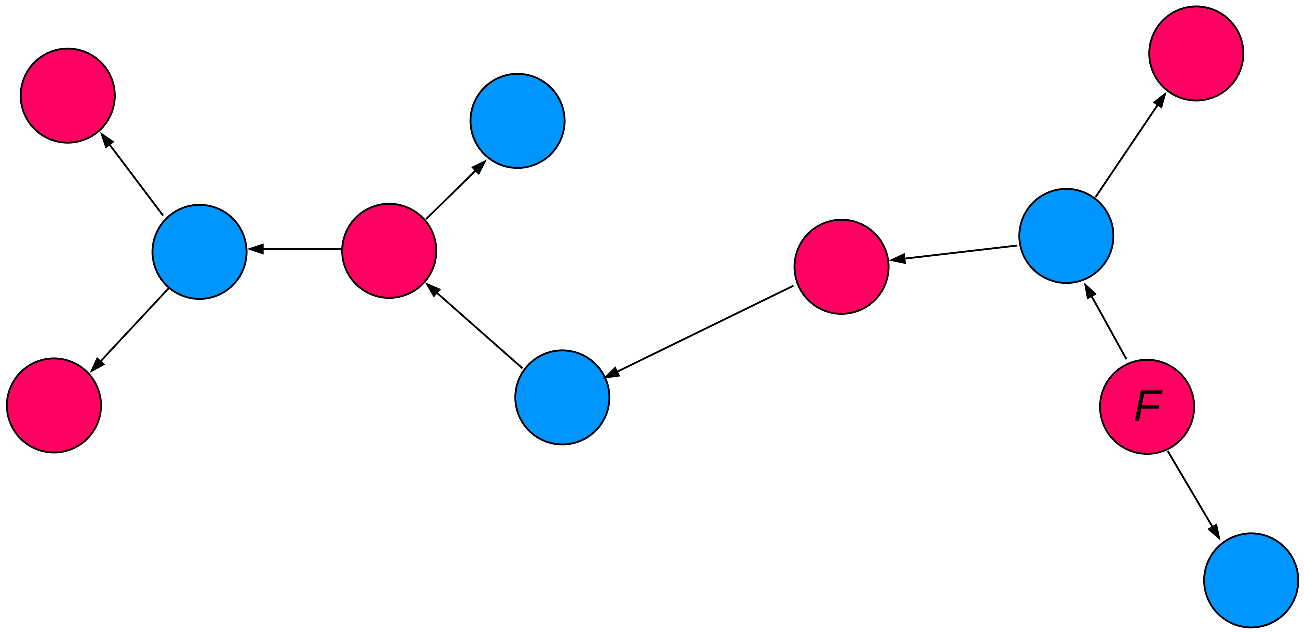
Cuckoo Graph

Edge Insertion: Good case 1: joining 2 trees



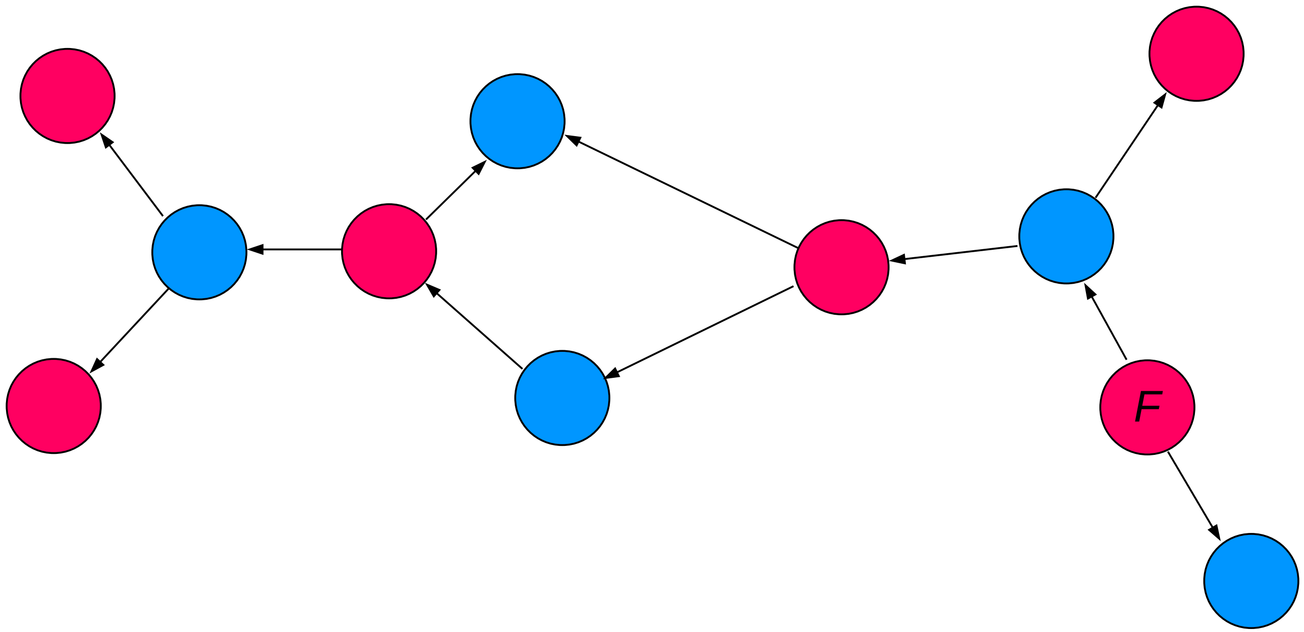
Cuckoo Graph

Edge Insertion: Good case 1: joining 2 trees



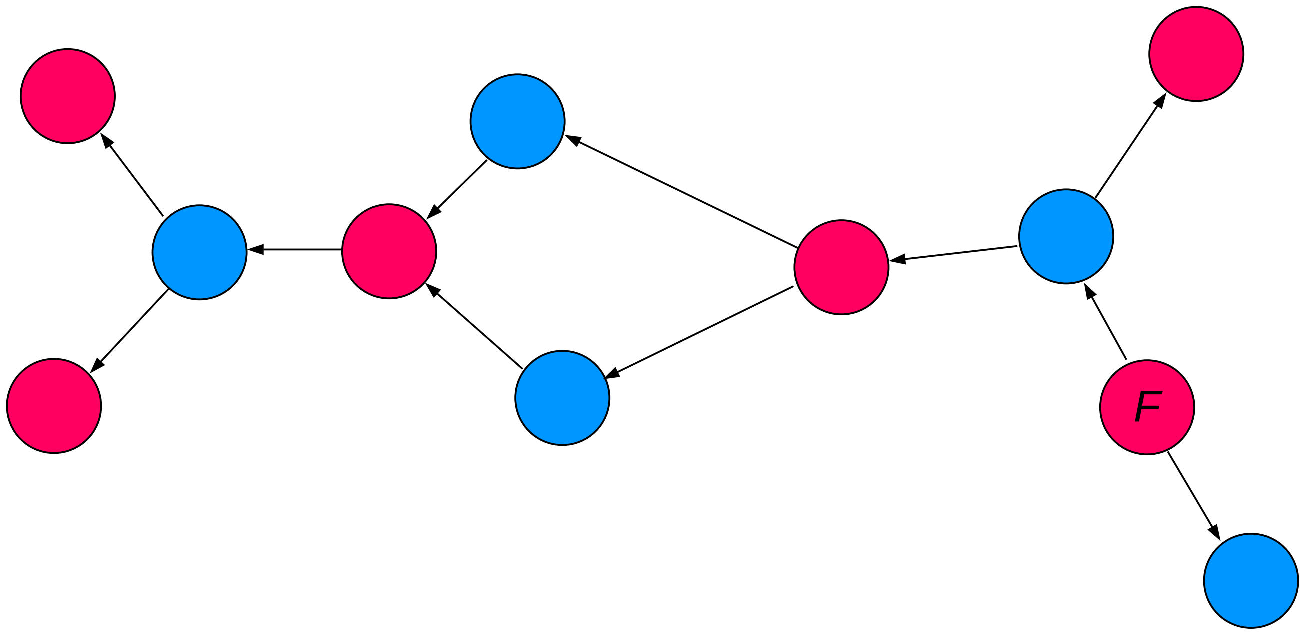
Cuckoo Graph

Edge Insertion: Good case 2: inserting an edge into a tree



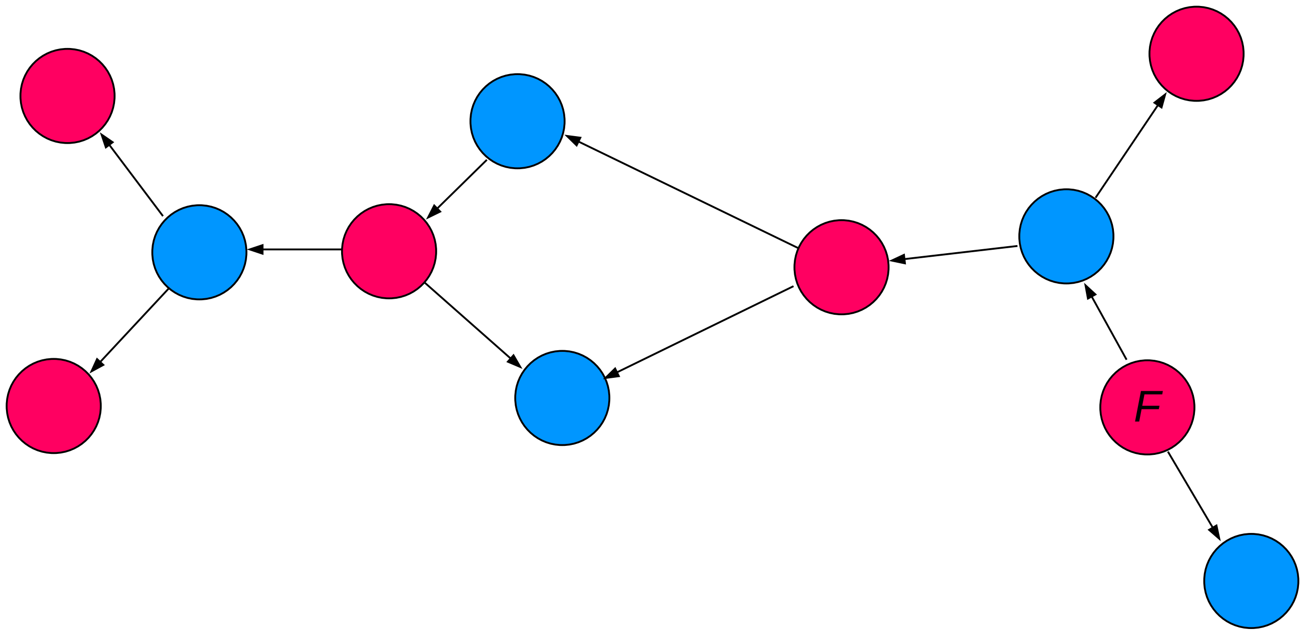
Cuckoo Graph

Edge Insertion: Good case 2: inserting an edge into a tree



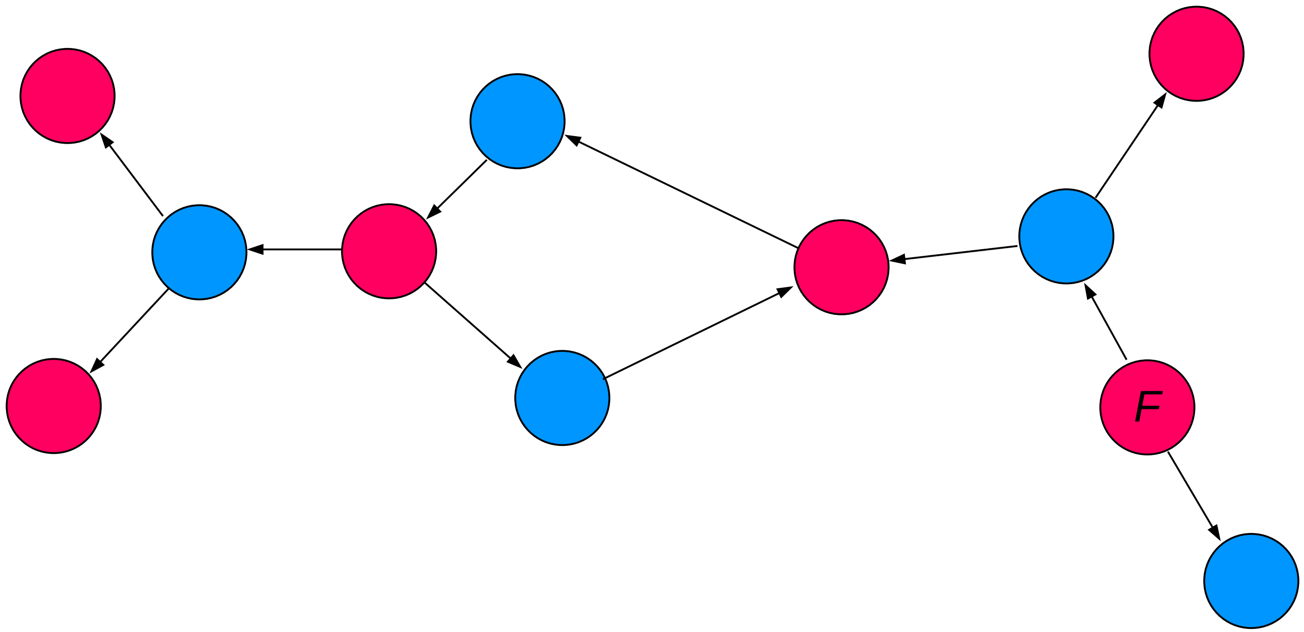
Cuckoo Graph

Edge Insertion: Good case 2: inserting an edge into a tree



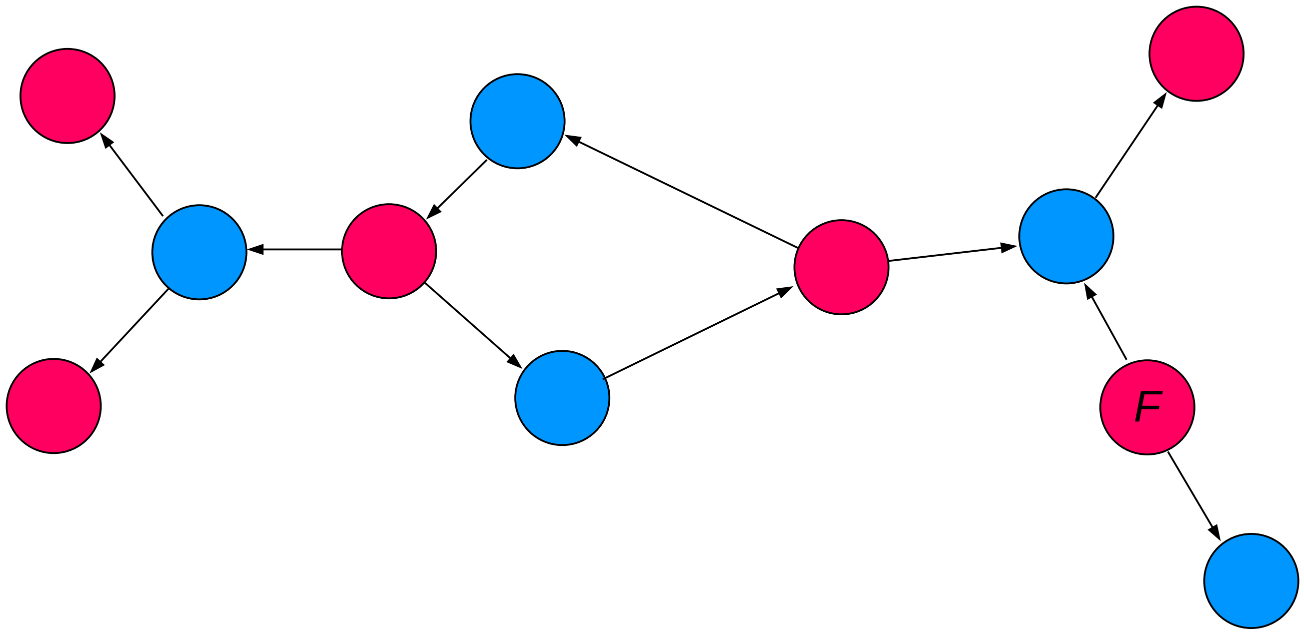
Cuckoo Graph

Edge Insertion: Good case 2: inserting an edge into a tree



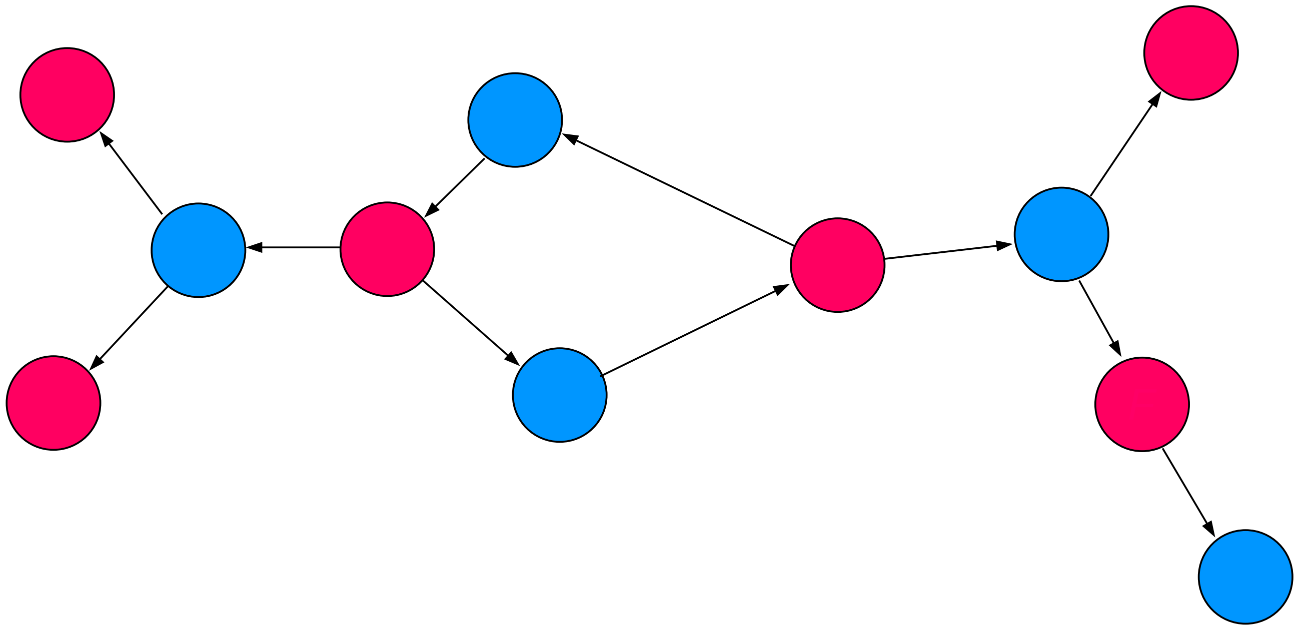
Cuckoo Graph

Edge Insertion: Good case 2: inserting an edge into a tree



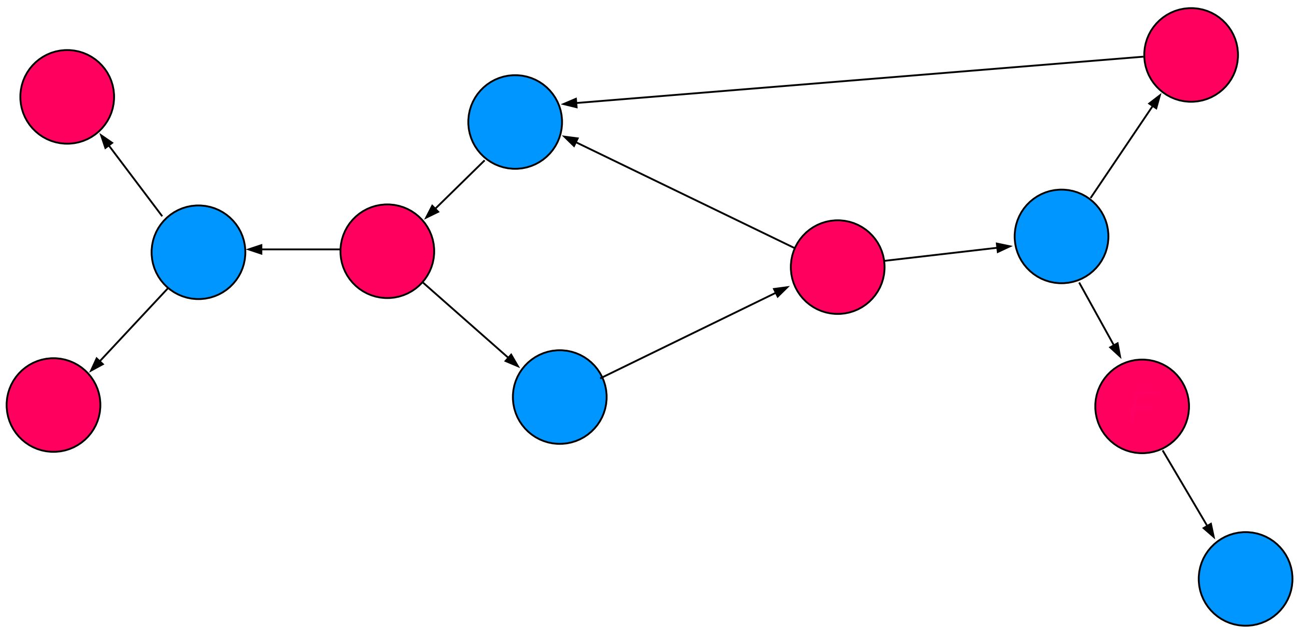
Cuckoo Graph

Edge Insertion: Good case 2: inserting an edge into a tree



Cuckoo Graph

Edge Insertion: Bad case: inserting an edge into a cyclic component



Cuckoo Hashing – Cuckoo Graph

	Cuckoo Hashing		Cuckoo Graph $G = (V_1, V_2, E)$
m	size of hash tables T_1, T_2	$\longleftrightarrow$	$\lvert V_1\rvert, \lvert V_2\rvert$
n	number of keys	$\longleftrightarrow$	$\lvert E\rvert$
	successful hashing	$\longleftrightarrow$	no complex components
	running time	$\longleftrightarrow$	size of components

Random Bipartite Graph

- $G_{m_1,m_2,n}$ random bipartite multigraph (labeled)
- $|V_1| = m_1$, $|V_2| = m_2$ labeled vertex sets
- $|E| = n$ labeled multi edges
- Each of these $\boxed{\#G_{m_1,m_2,n} = m_1^n m_2^n}$ graphs is **equally likely**

The edge labels encode the insertion procedure (**dynamic model**)

Notation: $G^{\circ}_{m_1,m_2,n}$ ... graphs with **no** complex component

Asymptotic Results

Probability of Successfull Hashing

Theorem 1

Suppose that $\boxed{n = (1-\varepsilon)m}$ for some $\varepsilon > 0$.

$$\Longrightarrow \quad \boxed{\boxed{\frac{\#G^{\circ}_{m,m,n}}{\#G_{m,m,n}} = 1 - \frac{h(\varepsilon)}{m} + O(m^{-2})}} \quad (m \to \infty)$$

with

$$\begin{aligned} h(\varepsilon) &= \frac{(2\varepsilon^2 - 5\varepsilon + 5)(1-\varepsilon)^3}{12(2-\varepsilon)^2\varepsilon^3} \\ &= \frac{5}{48}\varepsilon^{-3} - \frac{5}{16}\varepsilon^{-2} + \frac{21}{64}\varepsilon^{-1} - \frac{13}{96} + \frac{3}{256}\varepsilon + \frac{1}{256}\varepsilon^2 + \frac{1}{1024}\varepsilon^3 + O\left(\varepsilon^4\right) \end{aligned}$$

Asymptotic Results

Remark 1. [Devroye and Morin]: $1 - \frac{\#G^{\circ}_{m,m,n}}{\#G_{m,m,n}} = O(1/m)$.

Remark 2. The probability that Cuckoo hashing **fails** (with table sizes m and $n = (1-\varepsilon)m$ keys) is

$$\frac{h(\varepsilon)}{m} + O(m^{-2}).$$

Asymptotic Results

Probability of Successfull Hashing

Theorem 2

Suppose that $n = m$.

$$\Longrightarrow \quad \boxed{\frac{\#G^{\circ}_{m,m,n}}{\#G_{m,m,n}} = \sqrt{\frac{2}{3}} + o(1) = 0.8164965809\ldots + o(1)} \quad (m \to \infty)$$

Remark 3. The same results holds for *ususal* random graphs [Flajolet, Knuth, and Pittel, 1989]

Remark 4. Threshold appears at $\boxed{n = m - \Theta(m^{2/3})}$ (as for random graphs – **birth of a giant component**).

Asymptotic Results

Unicyclic Components

Theorem 3. Let $X_{m,n}$ denote the number of points in unicyclic components. Suppose that $\boxed{\varepsilon > 0}$ and $\boxed{n = (1-\varepsilon)m}$. Then, as $m \to \infty$, $X_{m,n}$ has a discrete limiting distribution with expected value

$$\boxed{\mathbf{E}\, X_{m,n} = \frac{(1-\varepsilon)^2}{(2-\varepsilon)\varepsilon^2} + O\left(\frac{1}{m}\right)}$$

and variance

$$\boxed{\mathbf{Var}\, X_{m,n} = \frac{(1-\varepsilon)^2(\varepsilon^2 - 3\varepsilon + 4)}{(2-\varepsilon)^2\varepsilon^4} + O\left(\frac{1}{m}\right)}.$$

Asymptotic Results

Tree Sizes

Theorem 4. Let $T_{k;m,n}$ denote the number of trees of size k. Suppose that $\boxed{\varepsilon > 0}$ and $\boxed{n = (1-\varepsilon)m}$. Then, as $m \to \infty$, $T_{k;m,n}$ satisfies a central limit theorem with expected value

$$\boxed{\mathbf{E}\, T_{k;m,n} = 2m \frac{k^{k-2}(1-\varepsilon)^{k-1} e^{k(\varepsilon-1)}}{k!} + O(1)}$$

and variance

$$\boxed{\mathbf{Var}\, T_{k;m,n} = 2m \left(\frac{k^{k-2}(1-\varepsilon)^{k-1} e^{k(\varepsilon-1)}}{k!} - \frac{k^{2k-4}(1-\varepsilon)^{2k-3} e^{2k(\varepsilon-1)} (k^2\varepsilon^2 + k^2\varepsilon - 4k\varepsilon + 2)}{(k!)^2} \right) + O(}$$

Asymptotic Results

Remark 1. Cyclic components are negligible (for $\varepsilon > 0$)

Remark 2. Expected average tree size ≤ 2
$\Longrightarrow$ expected running time $\leq 2m$

(This explains the extremely good performance of Cuckoo hashing.)

Generating Functions

Bipartite Trees

t_{1,m_1,m_2} ... number of bipartite **rooted** trees with m_1 nodes of type 1, m_2 nodes of type 2, and the root is of type 1.

t_{2,m_1,m_2} ... number of bipartite **rooted** trees with m_1 nodes of type 1, m_2 nodes of type 2, and the root is of type 2.

$$t_1(x,y) = \sum_{m_1,m_2} t_{1,m_1,m_2} \frac{x^{m_1}}{m_1!} \frac{y^{m_2}}{m_2!}, \qquad t_2(x,y) = \sum_{m_1,m_2} t_{2,m_1,m_2} \frac{x^{m_1}}{m_1!} \frac{y^{m_2}}{m_2!},$$

$$\boxed{t_1(x,y) = x e^{t_2(x,y)}, \quad t_2(x,y) = y e^{t_1(x,y)}}$$

Generating Functions

Bipartite Trees

$\tilde{t}_{m_1,m_2}$ ... number of bipartite **unrooted** trees with m_1 nodes of type 1 and m_2 nodes of type 2.

$$\tilde{t}(x,y) = \sum_{m_1,m_2} \tilde{t}_{m_1,m_2} \frac{x^{m_1}}{m_1!} \frac{y^{m_2}}{m_2!},$$

$$\boxed{\tilde{t}(x,y) = t_1(x,y) + t_2(x,y) - t_1(x,y)t_2(x,y)}$$

Lemma. [Scoins, 1962]

$$\boxed{\tilde{t}_{m_1,m_2} = m_1^{m_2-1} m_2^{m_1-1}}$$

Generating Functions

Bipartite Trees

Remark 1 $t_{1,m_1,m_2} = m_1 \tilde{t}_{m_1,m_2}$, $t_{2,m_1,m_2} = m_2 \tilde{t}_{m_1,m_2}$,

Remark 2 $t_1(x,x) = t_2(x,x) = t(x)$ is the usual tree function given by

$$t(x) = xe^{t(x)}.$$

Remark 3 $t_{m_1,m_2,n}$ ... number of unrooted labeled bipartite trees with m_1 nodes of type 1, m_2 nodes of type 2, and n (labeled) edges.

$$\sum_{m_1,m_2,n} t_{m_1,m_2,n} \frac{x^{m_1}}{m_1!} \frac{y^{m_2}}{m_2!} \frac{u^n}{n!} = \frac{1}{u} \tilde{t}(xu, yu).$$

Generating Functions

Lemma

$$g^{\circ}(x,y,u) = \sum_{m_1,m_2,n} \#G^{\circ}_{m_1,m_2,n} \frac{x^{m_1}}{m_1!}\frac{y^{m_2}}{m_2!}\frac{u^n}{n!} = \frac{e^{\frac{1}{u}\tilde{t}(xu,yu)}}{\sqrt{1 - t_1(xu,yu)t_2(xu,yu)}}.$$

Proof.

Cyclic component with $2k$ cyklic points: $\frac{1}{2k} t_1(xu,yu)^k t_2(xu,yu)^k$

$$\Longrightarrow \quad g^{\circ}(x,y,u) = \exp\left(\frac{1}{u}\tilde{t}(xu,yu) + \sum_{k\geq 1} \frac{1}{2k} t_1(xu,yu)^k t_2(xu,yu)^k\right)$$

Generating Functions

Corollary

$$
\begin{aligned}
\#G^{\circ}_{m_1,m_2,n} &= \frac{m_1!m_2!n!}{(m_1+m_2-n)!}[x^{m_1}y^{m_2}]\frac{\tilde{t}(x,y)^{m_1+m_2-n}}{\sqrt{1-t_1(x,y)t_2(x,y)}} \\
&= -\frac{m_1!m_2!n!}{4\pi(m_1+m_2-n)!} \\
&\qquad \times \boxed{\int_{|x|=x_0}\int_{|y|=y_0}\frac{\tilde{t}(x,y)^{m_1+m_2-n}}{\sqrt{1-t_1(x,y)t_2(x,y)}}\frac{dx}{x^{m_1+1}}\frac{dy}{y^{m_2+1}}}.
\end{aligned}
$$

$\longrightarrow$ DOUBLE SADDLE POINT

Generating Functions

Extensions

E.g., in

$$\frac{e^{\frac{1}{u}\tilde{t}(xu,yu)+x(w-1)+y(w-1)}}{\sqrt{1-t_1(xu,yu)t_2(xu,yu)}}.$$

the additional variable w counts the number of **isolated nodes** (= tree components of size 1).

etc.

Double Saddle Point

Lemma

$f(x,y)$, $g(x,y)$ ... analytic functions around $(0,0)$
(+ minor technical assumptions)

$$\Longrightarrow \boxed{[x^{m_1}y^{m_2}]g(x,y)f(x,y)^k = \frac{g(x_0,y_0)f(x_0,y_0)^k}{2\pi x_0^{m_1}y_0^{m_2}k\sqrt{\Delta}}\left(1+\frac{h}{24\Delta^3}\frac{1}{k}+O\left(\frac{1}{k^2}\right)\right)},$$

where x_0 and y_0 are uniquely defined by

$$\frac{m_1}{k} = \frac{x_0}{f(x_0,y_0)}\left[\frac{\partial}{\partial x}f(x,y)\right]_{(x_0,y_0)}, \qquad \frac{m_2}{k} = \frac{y_0}{f(x_0,y_0)}\left[\frac{\partial}{\partial y}f(x,y)\right]_{(x_0,y_0)}$$

and are contained in a finite interval of the positive real line, that is, m_1, m_2, and k have to be of the same order of magnitude.

Double Saddle Point

Set

$$\kappa_{ij} = \left[\frac{\partial^i}{\partial u^i}\frac{\partial^j}{\partial v^j}\log f(x_0 e^u, y_0 e^v)\right]_{(0,0)}, \quad \overline{\kappa}_{ij} = \left[\frac{\partial^i}{\partial u^i}\frac{\partial^j}{\partial v^j}\log g(x_0 e^u, y_0 e^v)\right]_{(0,0)}.$$

Then $\boxed{\Delta = \kappa_{20}\kappa_{02} - \kappa_{11}^2}$, and with

$$\begin{aligned}
\alpha &= 54\kappa_{21}\kappa_{11}\kappa_{12}\kappa_{20}\kappa_{02} + 6\kappa_{22}\kappa_{20}\kappa_{02}\kappa_{11}^2 - 12\kappa_{22}\kappa_{11}^4 + 4\kappa_{03}\kappa_{11}^3\kappa_{30} \\
&\quad + 36\kappa_{21}\kappa_{11}^3\kappa_{12} + 6\kappa_{22}\kappa_{20}^2\kappa_{02}^2 + 6\kappa_{03}\kappa_{11}\kappa_{30}\kappa_{20}\kappa_{02}, \\
\beta &= -5\kappa_{02}^3\kappa_{30}^2 + 30\kappa_{02}^2\kappa_{30}\kappa_{11}\kappa_{21} - 24\kappa_{02}\kappa_{30}\kappa_{12}\kappa_{11}^2 - 6\kappa_{02}^2\kappa_{30}\kappa_{12}\kappa_{20} \\
&\quad - 12\kappa_{11}\kappa_{02}^2\kappa_{31}\kappa_{20} - 36\kappa_{02}\kappa_{21}^2\kappa_{11}^2 - 9\kappa_{02}^2\kappa_{21}^2\kappa_{20} + 3\kappa_{02}^3\kappa_{40}\kappa_{20} \\
&\quad - 3\kappa_{02}^2\kappa_{40}\kappa_{11}^2 + 12\kappa_{11}^3\kappa_{02}\kappa_{31}, \\
\gamma &= 12\Delta\left(\kappa_{02}^2\kappa_{30} - \kappa_{11}\kappa_{20}\kappa_{03} - 3\kappa_{21}\kappa_{11}\kappa_{02} + \kappa_{12}\kappa_{11}^2 + \kappa_{12}(\kappa_{02}\kappa_{20} + \kappa_{11}^2)\right), \\
\delta &= 24\Delta(\kappa_{11}\kappa_{20}\kappa_{02} - \kappa_{11}^3), \\
\eta &= 12\Delta(\kappa_{02}\kappa_{11}^2 - \kappa_{02}^2\kappa_{20})
\end{aligned}$$

we have

$$\boxed{h = \alpha + \beta + \widehat{\beta} + \gamma\overline{\kappa}_{10} + \widehat{\gamma}\overline{\kappa}_{01} + \delta\overline{\kappa}_{10}\overline{\kappa}_{01} + \eta\overline{\kappa}_{10}^2 + \widehat{\eta}\overline{\kappa}_{01}^2 + 4\eta\overline{\kappa}_{20} + 4\widehat{\eta}\overline{\kappa}_{02} + 4\delta\overline{\kappa}_{11}},$$

where $\widehat{\ }$ indicates to replace all functions of type κ_{ij} by κ_{ji}.

Double Saddle Point

Proof of Theorem 1

$m_1 = m_2 = m$, $n = (1-\varepsilon)m$ for some $\varepsilon > 0$

$$\longrightarrow \quad x_0 = y_0 = (1-\varepsilon)e^{\varepsilon-1} < \frac{1}{e}.$$

Application with $f(x,y) = \tilde{t}(x,y)$ and $k = 2m - n = (1+\varepsilon)n$

$$\longrightarrow \quad \text{Theorem 1.}$$

Double Saddle Point

Proof of Theorem 2

- Saddle point $x_0 = y_0 = \frac{1}{e}$ and squareroot singularity coincide !!!
- Apply Lagrange inversion for $t_1 = x \exp(ye^{t_1})$.
- Series expansion for $1/\sqrt{1-v}$
- Infinite series of double saddle point integrals
  (one integral with scale $\boxed{e^{-t^2/2}}$,
  the second integral with scale $\boxed{e^{-cs^3}}$)
- **Lommel functions** (similar to Airy functions)
- (Explicit) Mellin transform of Lommel function

$$\longrightarrow \sqrt{\frac{2}{3}} \qquad \text{(Theorem 2)}$$

Double Saddle Point

Lommel Functions

- Fundamental system of the inhomogeneous Bessel equation

$$x^2y'' + xy' - (x^2 + \nu^2) = kx^{\mu+1}.$$

- Closely related to the integral

$$\int_0^\infty e^{-t^3+kt} t \, dt.$$

Double Saddle Point

Remark

The *analytic structure* of generating functions for bipartite random graphs is more difficult than that of usual random graphs. This is due to the double saddle point (that comes from the additional variable) and the squareroot singularity that is now in 2 variables (instead of 1).

Nevertheless the results look the same. Thus, one can expect that most properties of random graphs have a counterpart in random bipartite graphs (birth of giant component etc.)

Thank You!