Using Pólya urns to show normal limit laws for fringe subtrees in m-ary search trees

Cecilia Holmgren, Uppsala University
Svante Janson and Matas Šileikis (Uppsala University)

Krakow, Poland
July 4, 2016

Aim of study

- To show *normal limit laws* for the number of *fringe subtrees* that are isomorphic to an arbitrary fixed tree T in the m-ary search tree with $m \leq 26$.

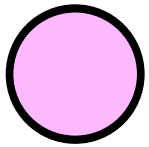
- To show *multivariate normal limit laws* for random vectors of such numbers for different fringe subtrees in the m-ary search tree.

The m-ary search tree

- Start with a tree containing just an empty root and the set $\{1, 2, \ldots, n\}$.
- Numbers/keys from the set are drawn at random, until it is exhausted. The first $m - 1$ keys are put in the root, and are placed in increasing order from left to right; they divide the set of real numbers into m intervals.
- When the root is full (after the first $m - 1$ keys are added), it gets m children that are initially empty, and each further key is passed to one of the children depending on which interval it belongs to.

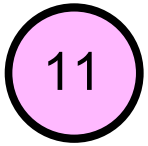
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



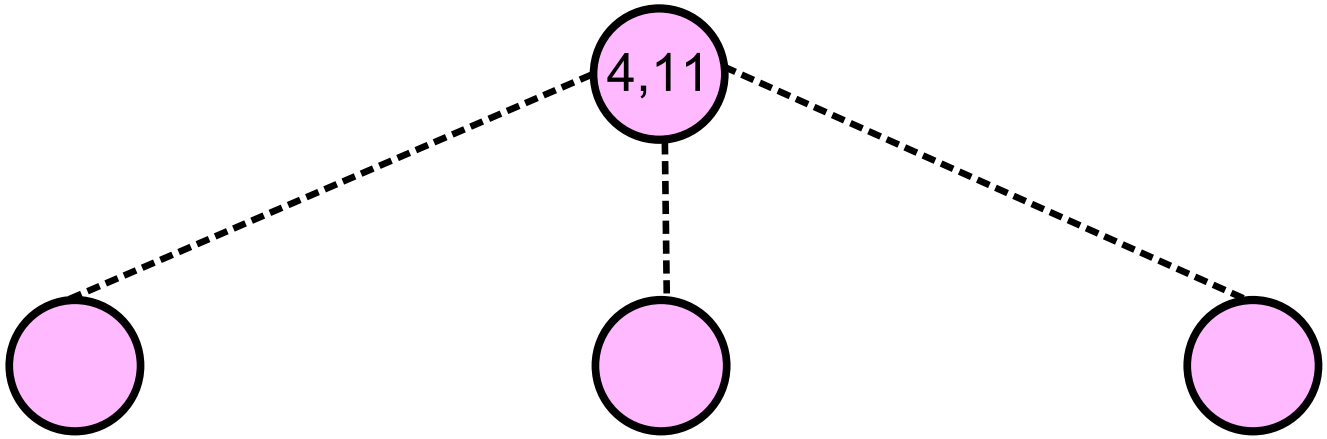
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



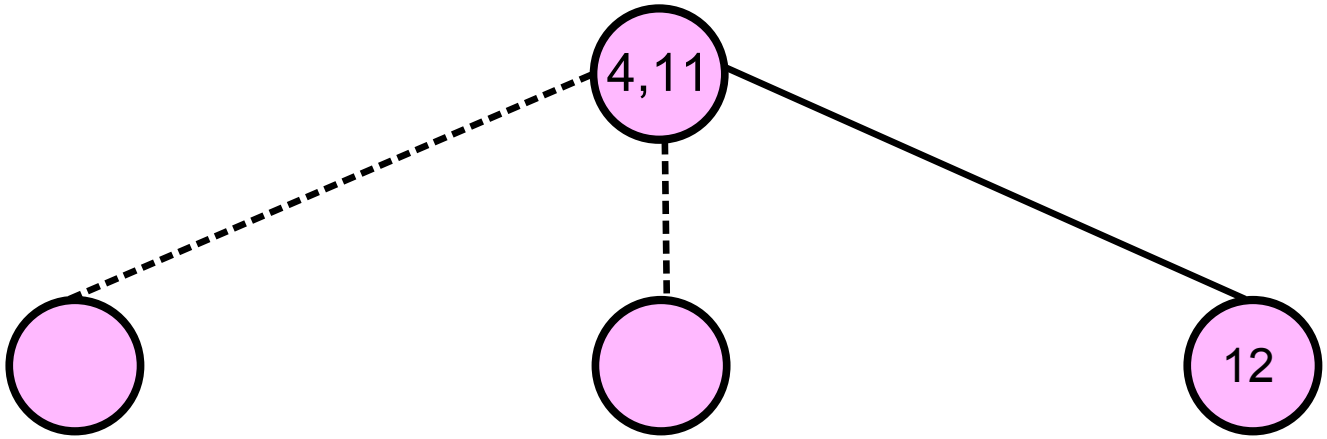
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



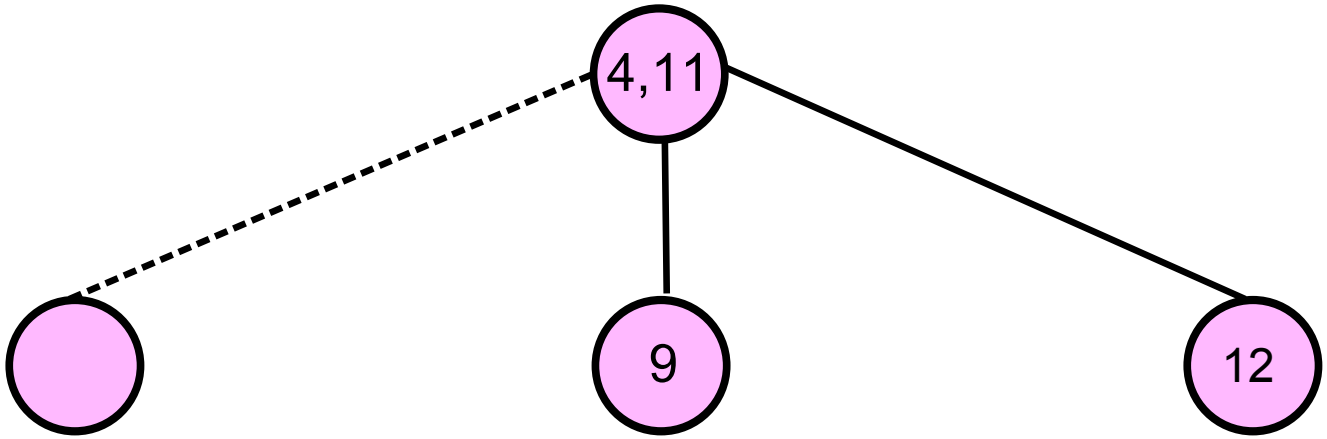
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



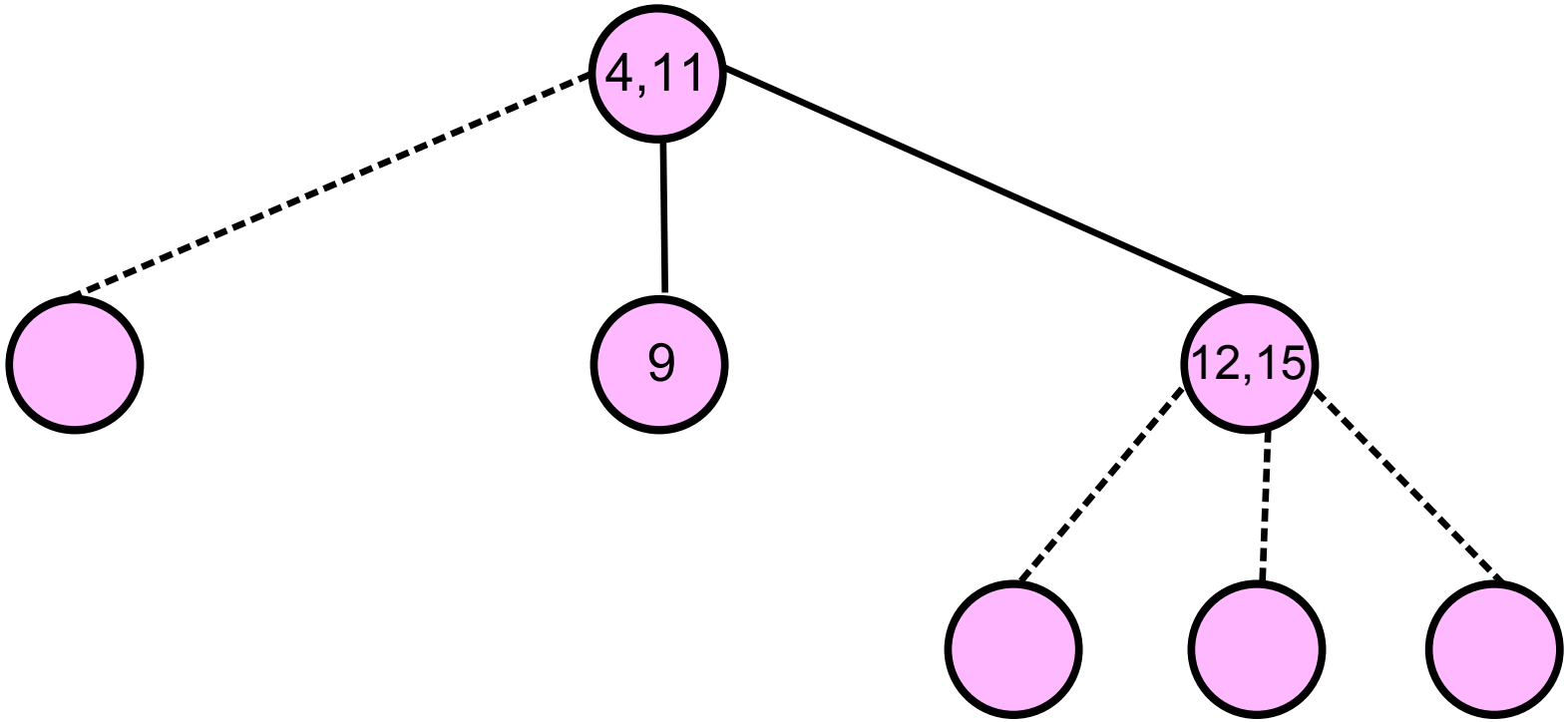
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



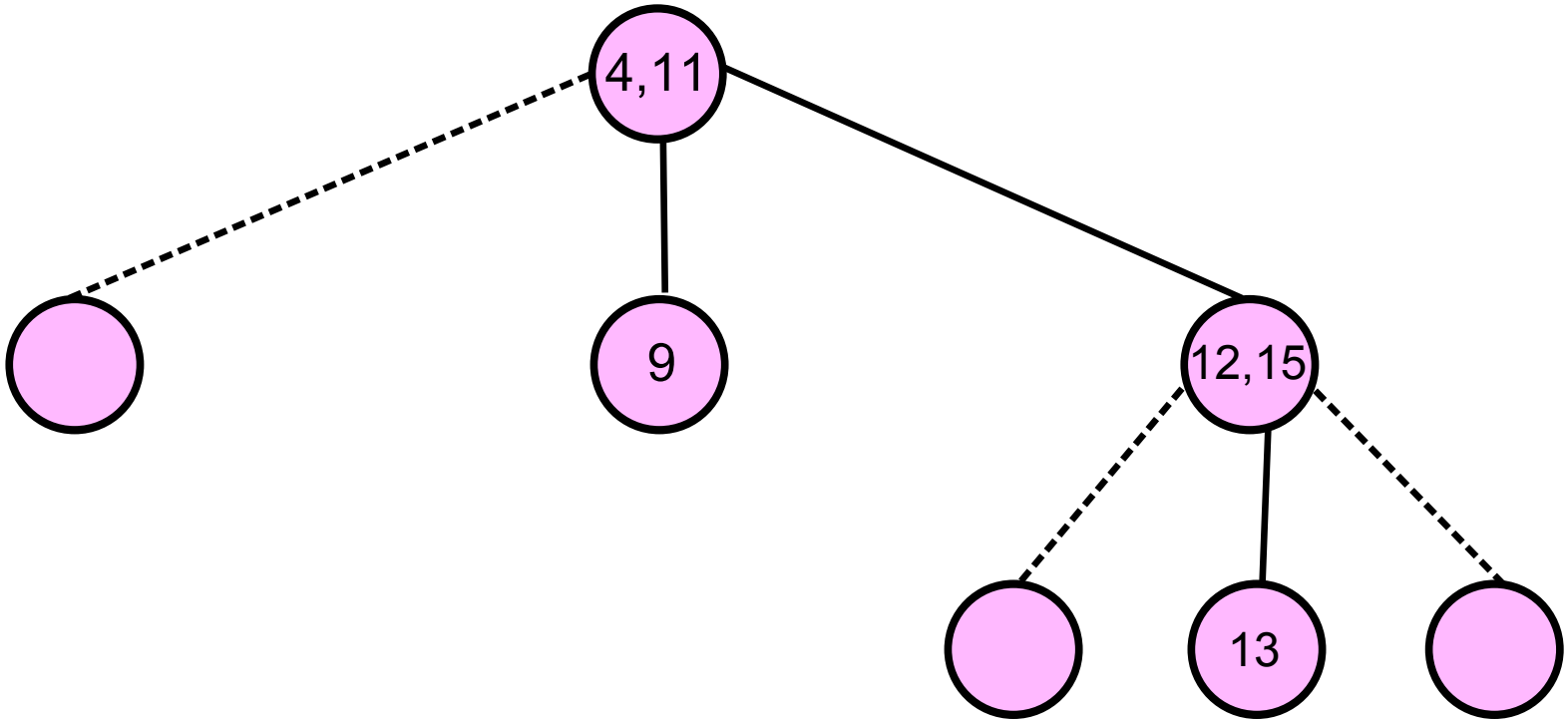
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



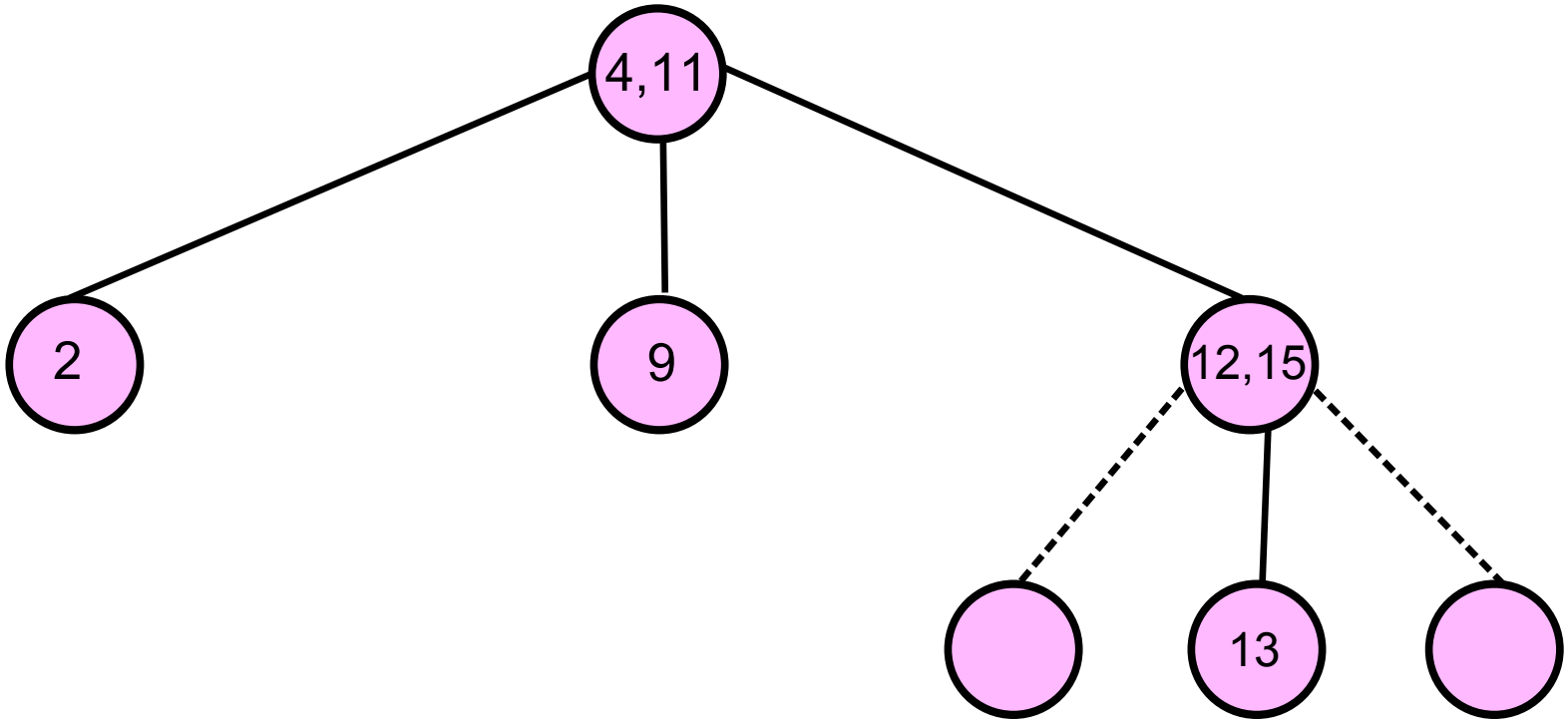
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



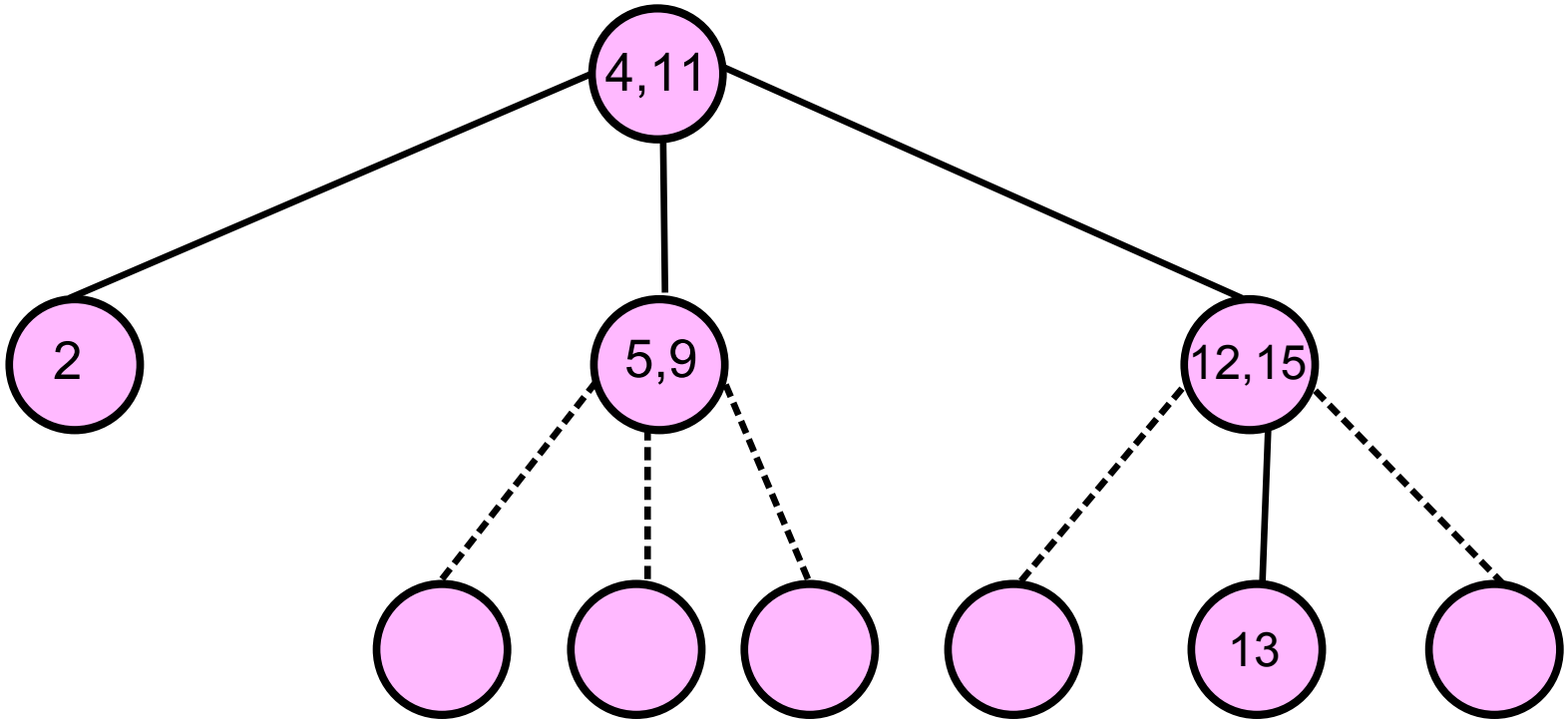
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



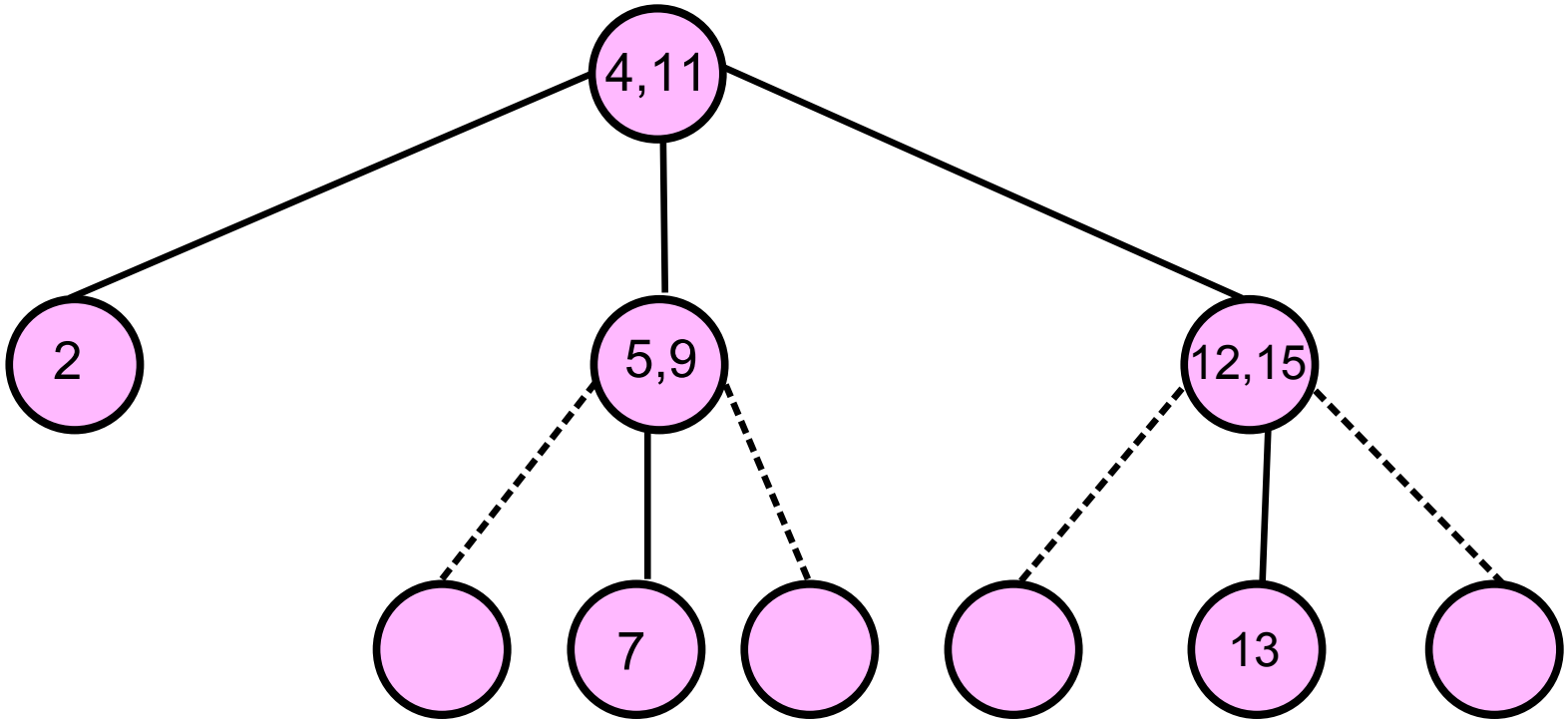
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



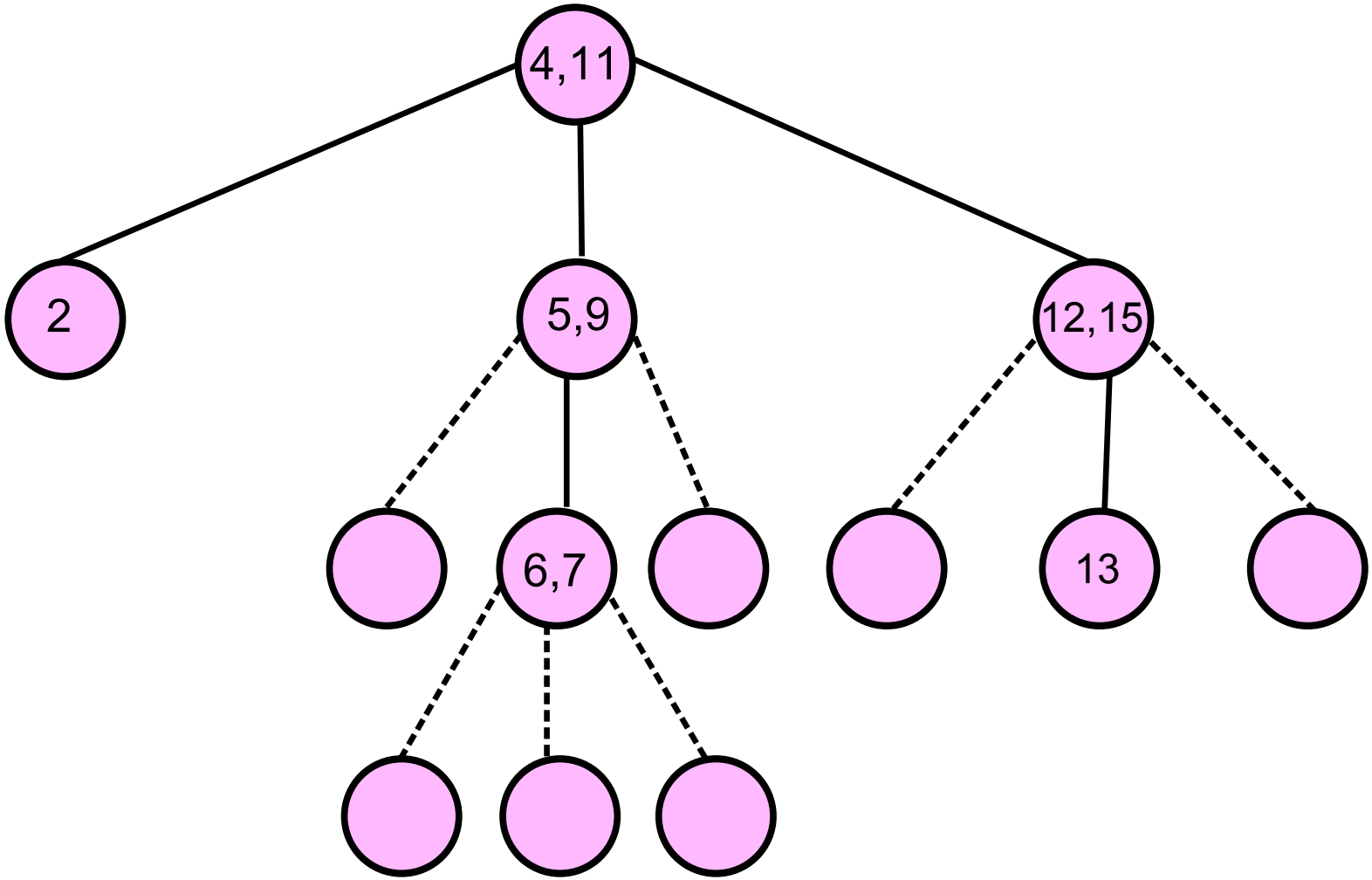
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



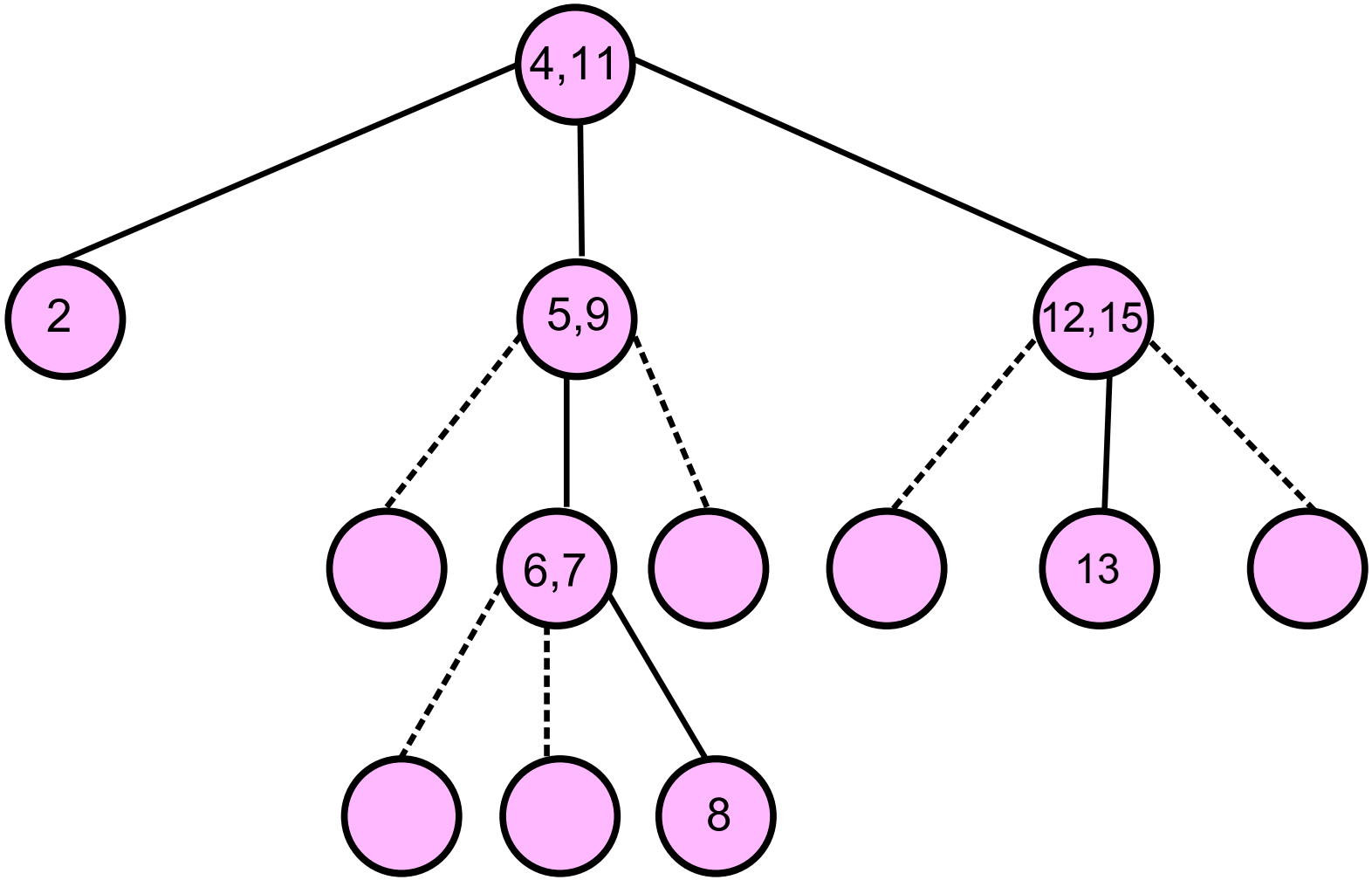
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



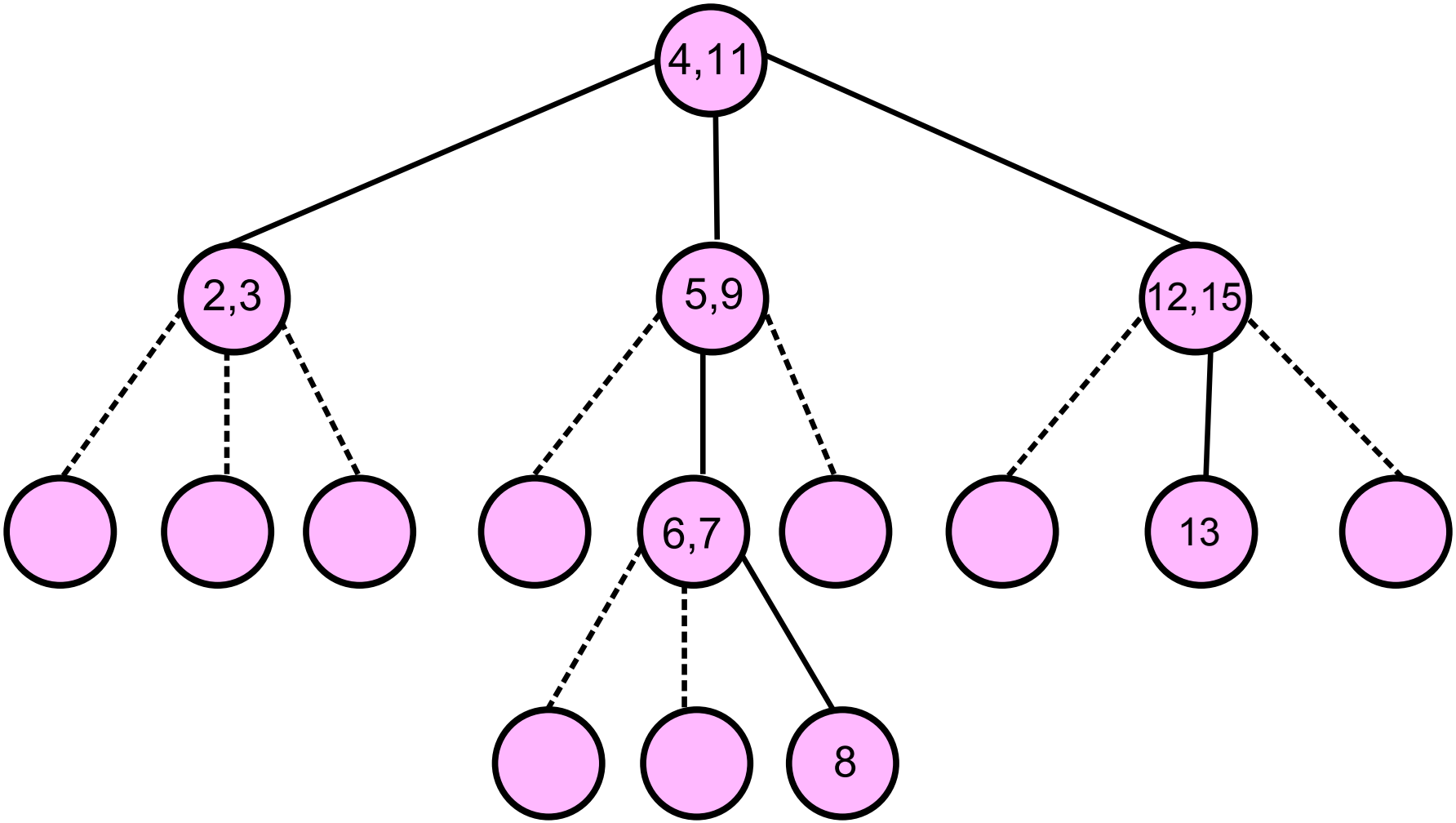
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



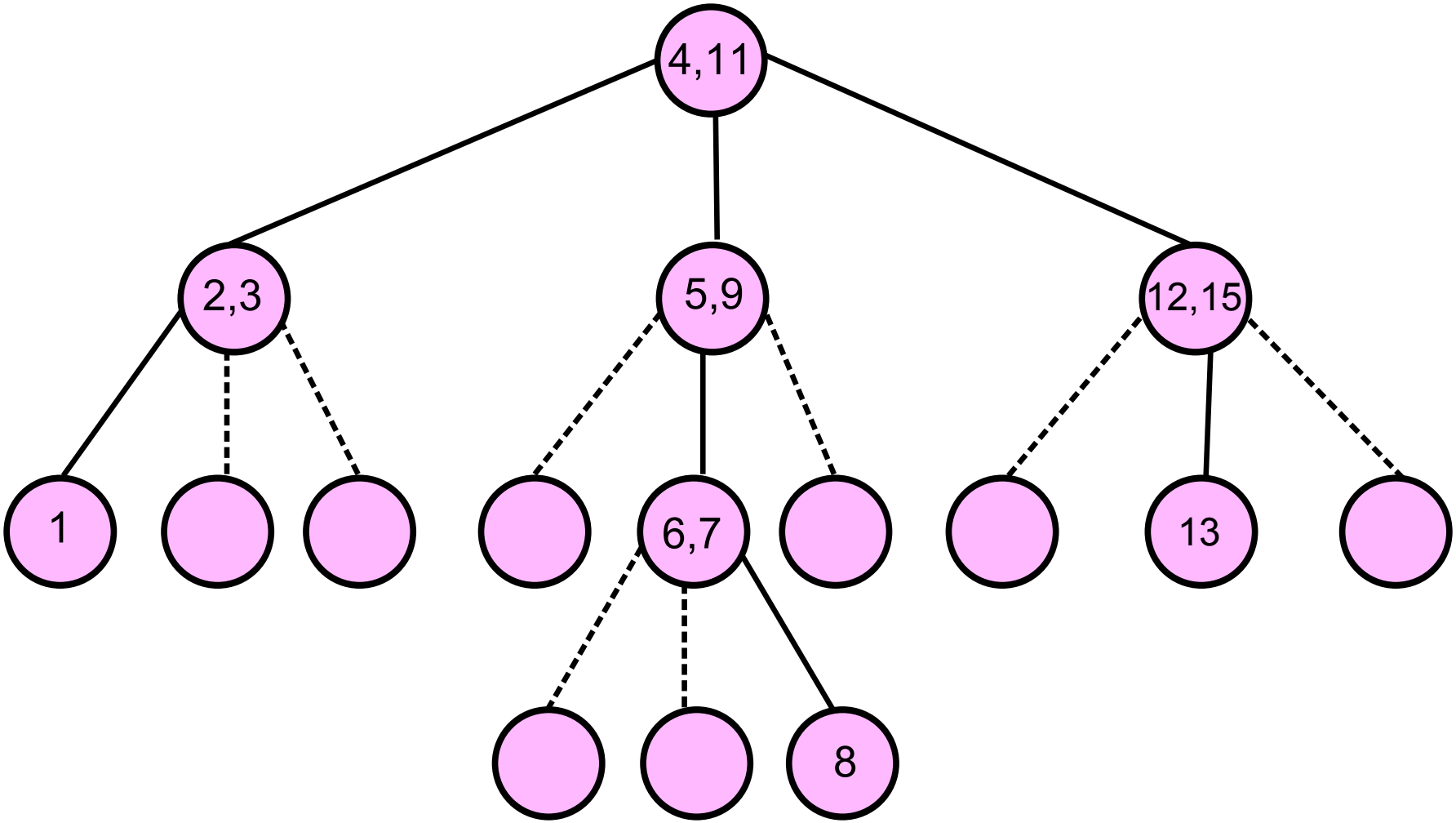
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



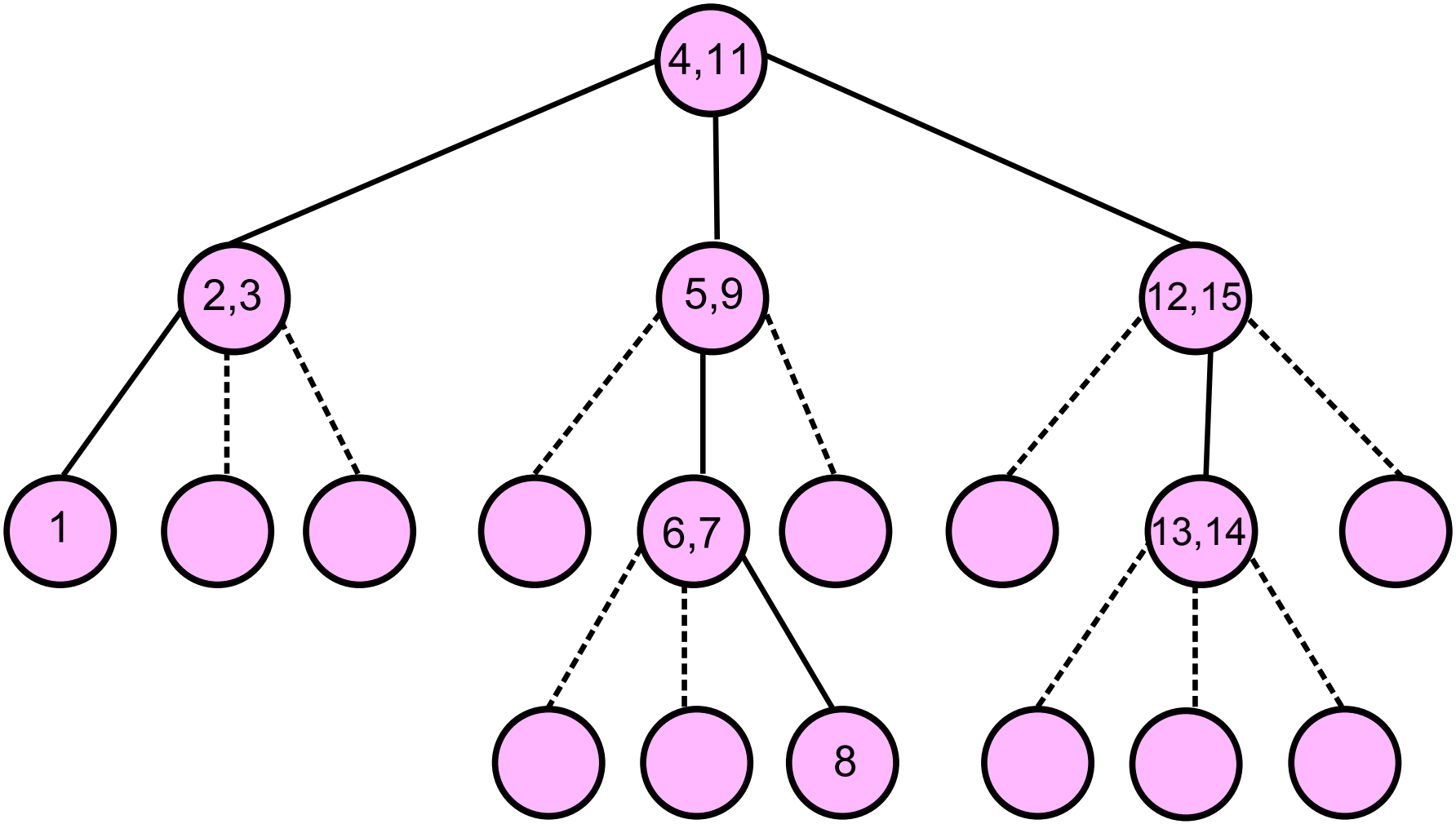
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



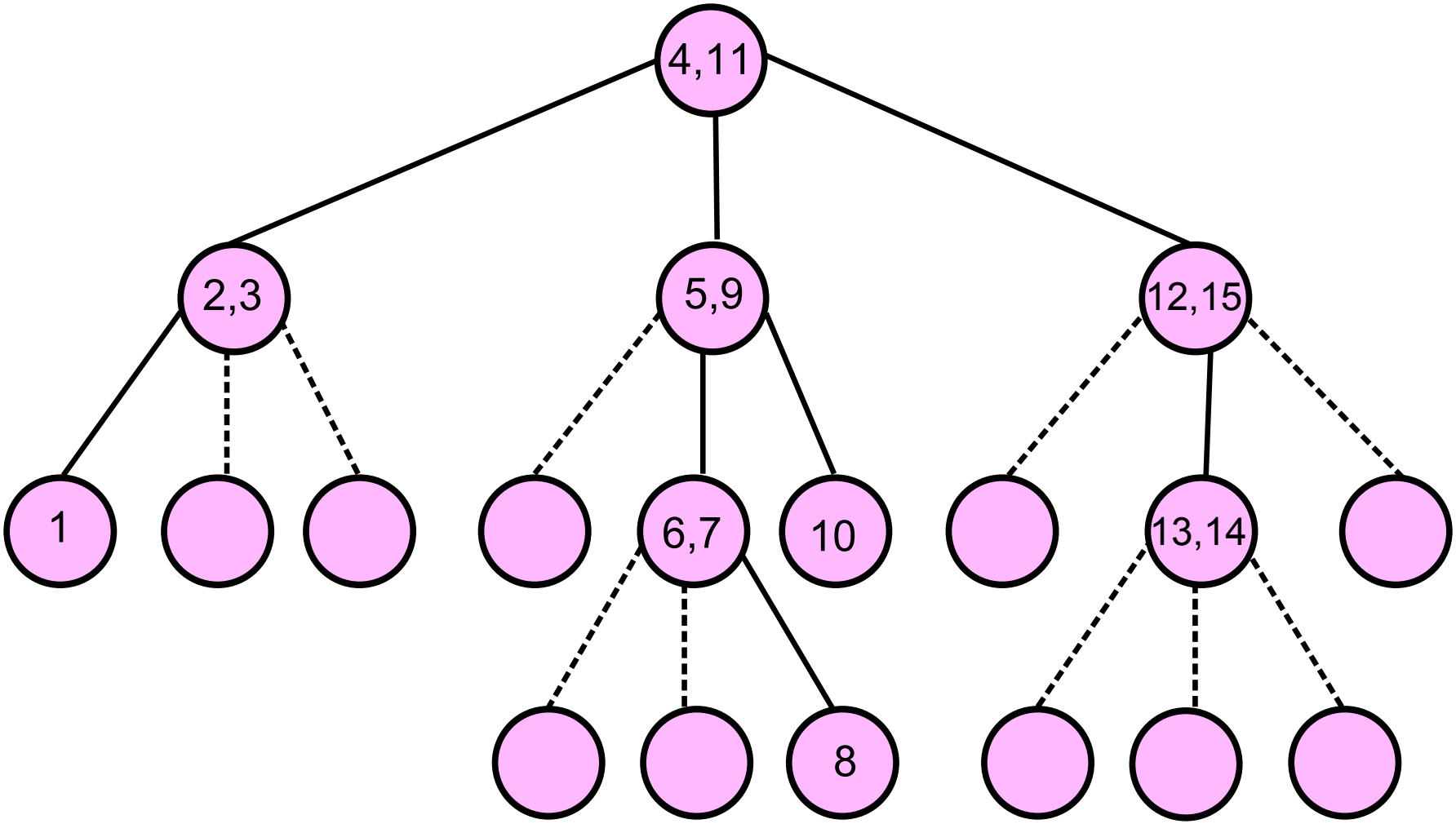
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



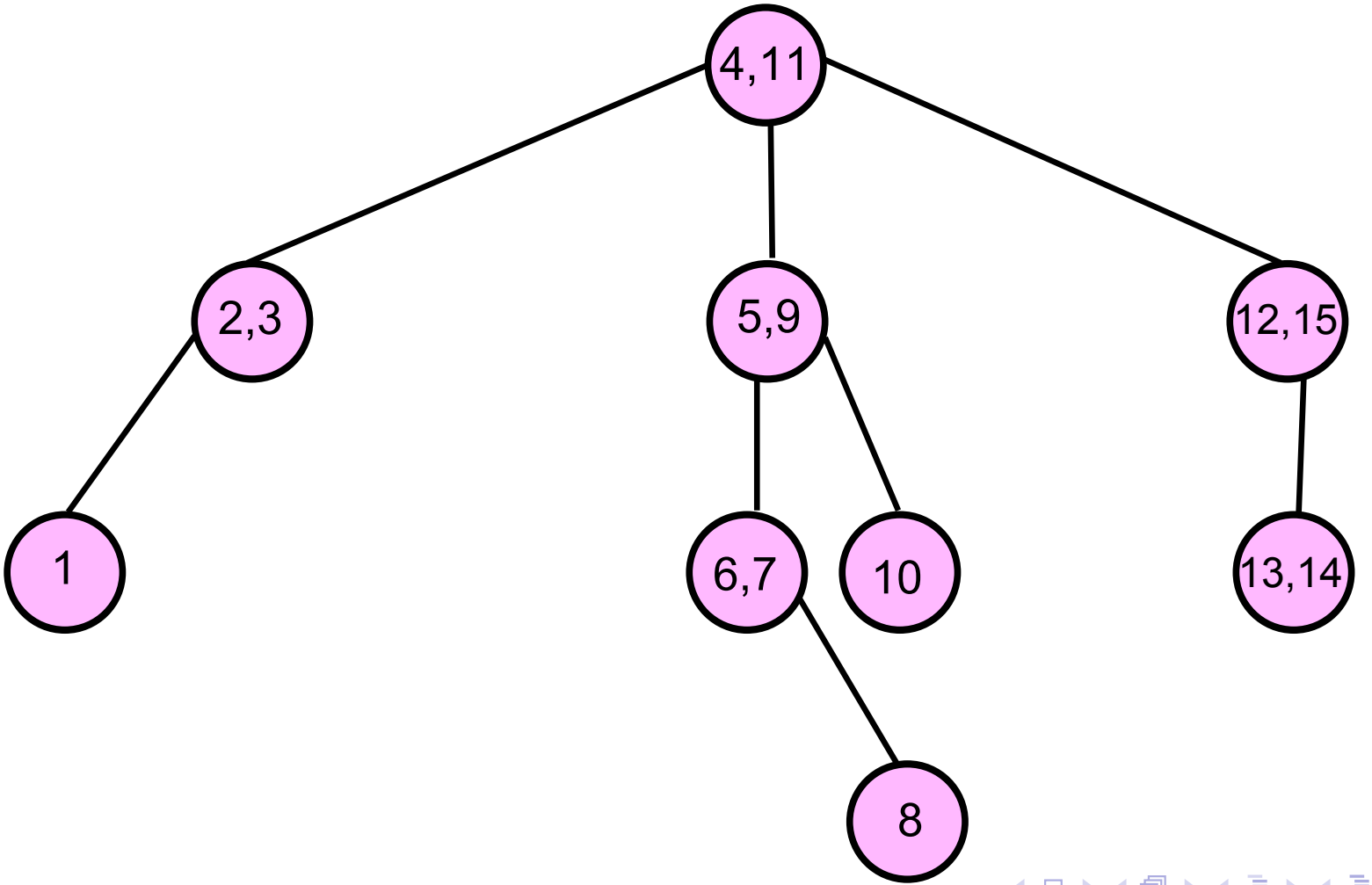
Example of a ternary search tree ($m = 3$)

We construct a ternary search tree from the set $\{1, 2, \ldots, 15\}$.



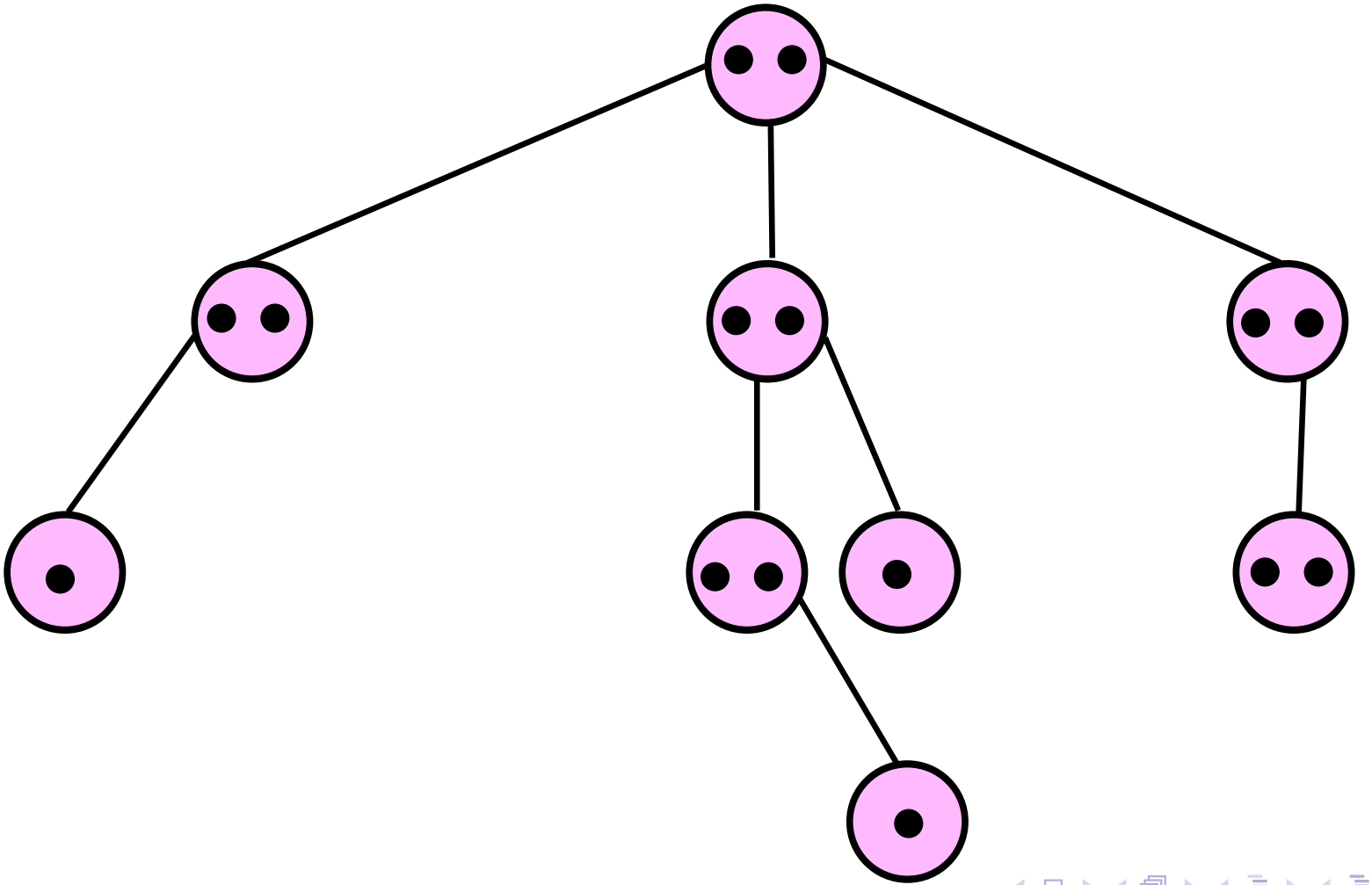
Example of a ternary search tree ($m = 3$)

Only the "*internal*" nodes, i.e., those that hold keys, are counted as part of the tree. The "*external*" (empty) nodes are ignored.



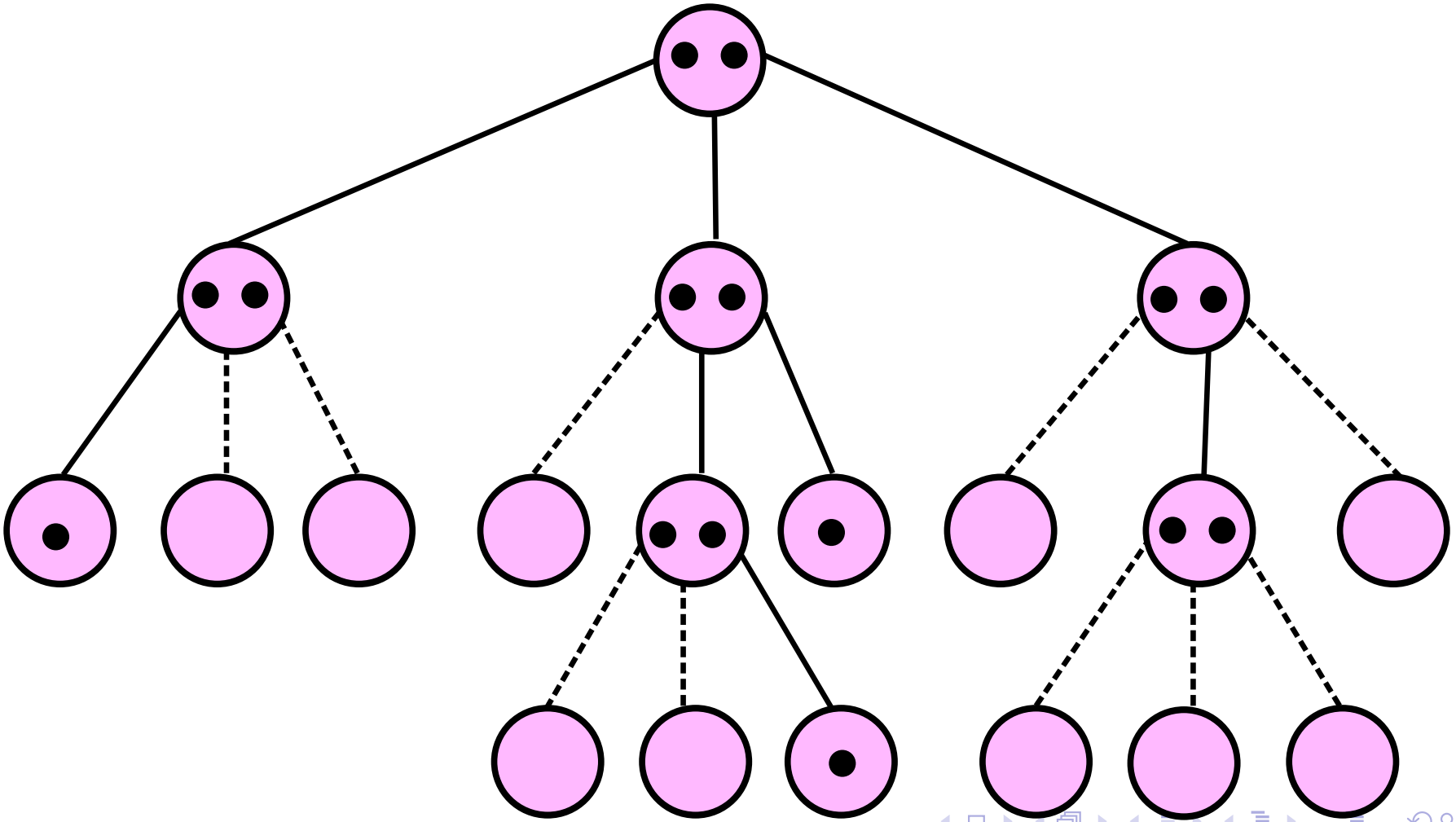
Example of a ternary search tree ($m = 3$)

Only the order relations of the keys are important, thus the keys can be replaced by identical items.



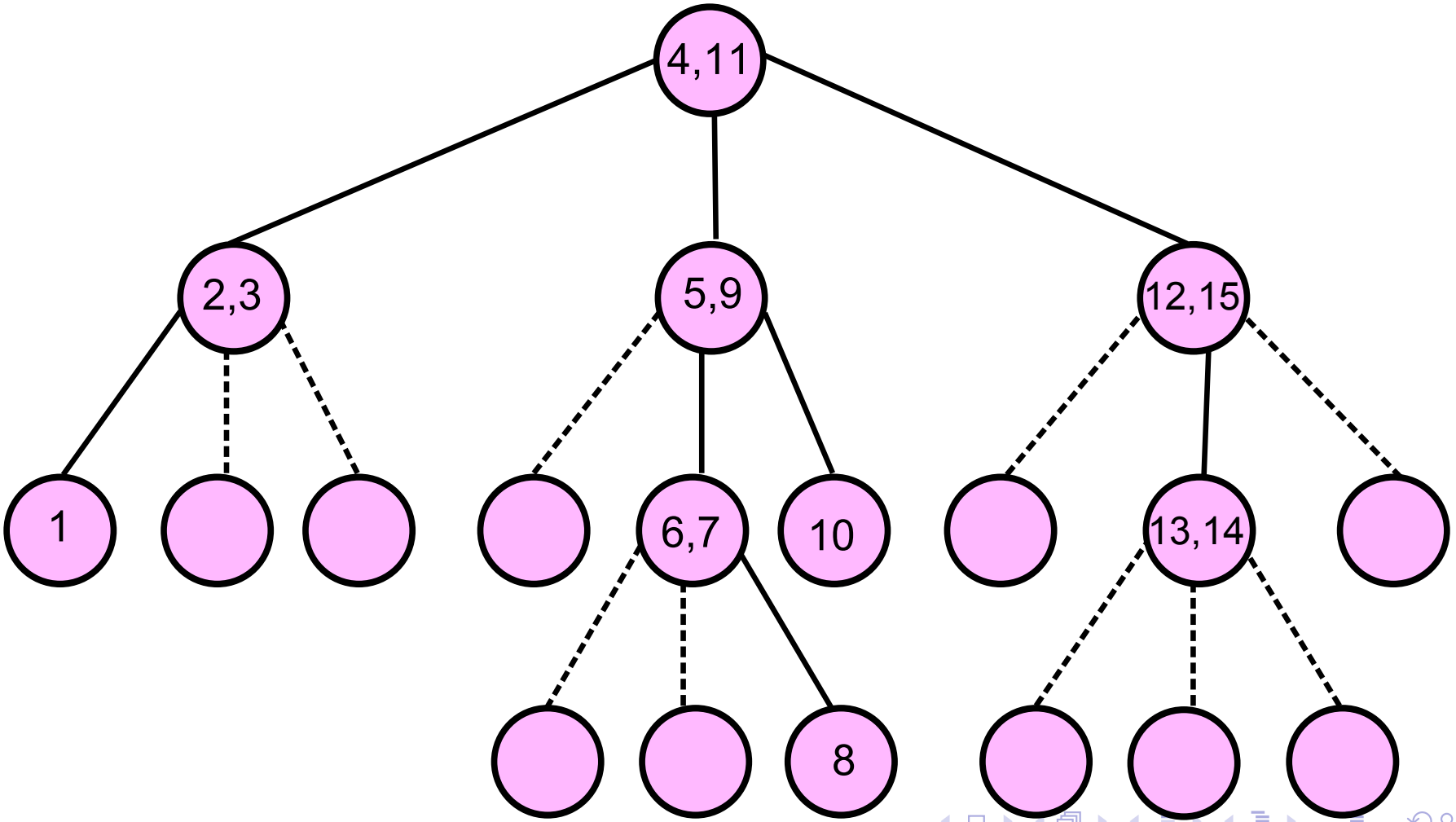
Example of a ternary search tree ($m = 3$)

An m-ary search tree with n keys/items has $n + 1$ possible places "*gaps*" with equal probability for insertion of a new key.

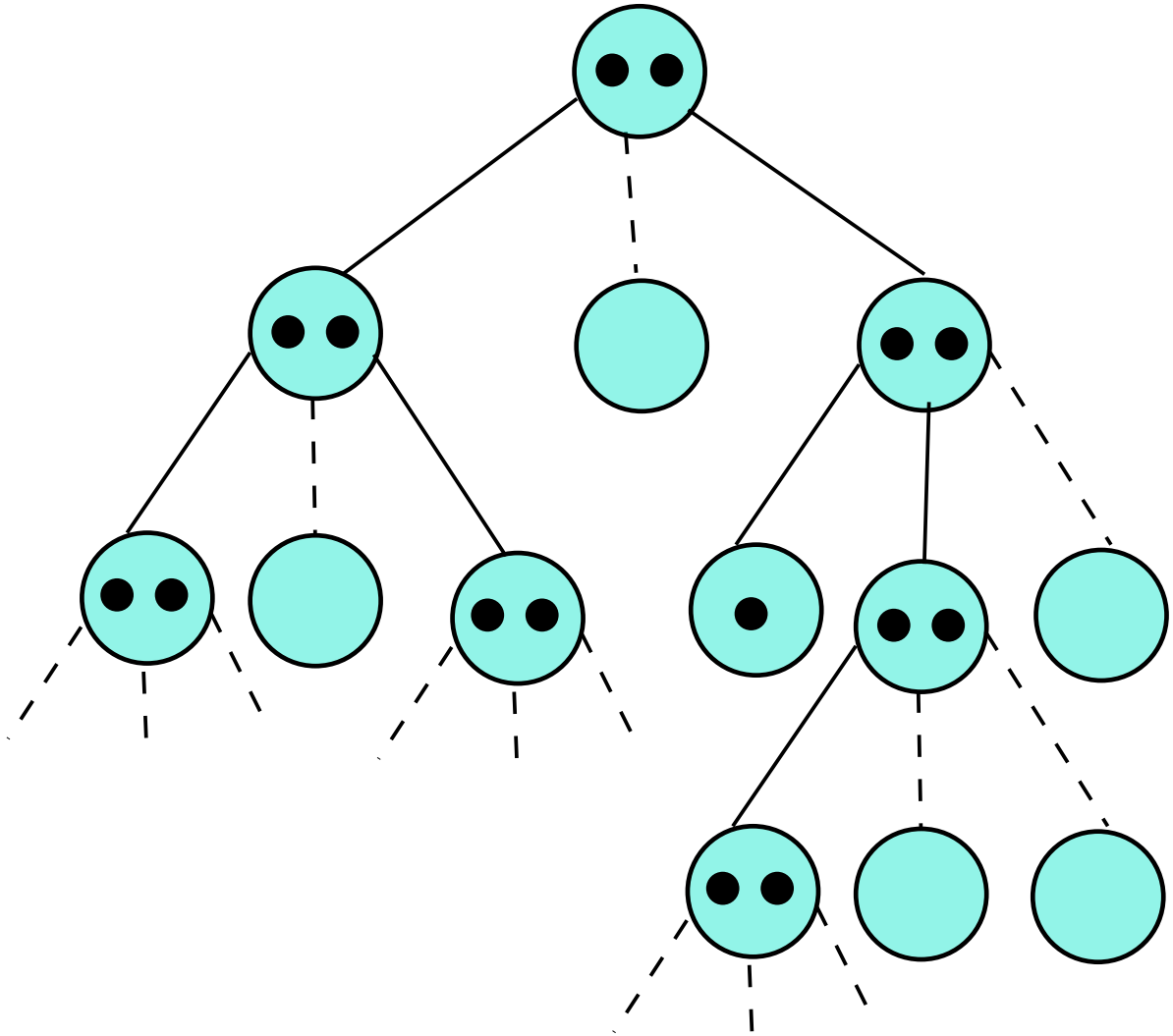


Example of a ternary search tree ($m = 3$)

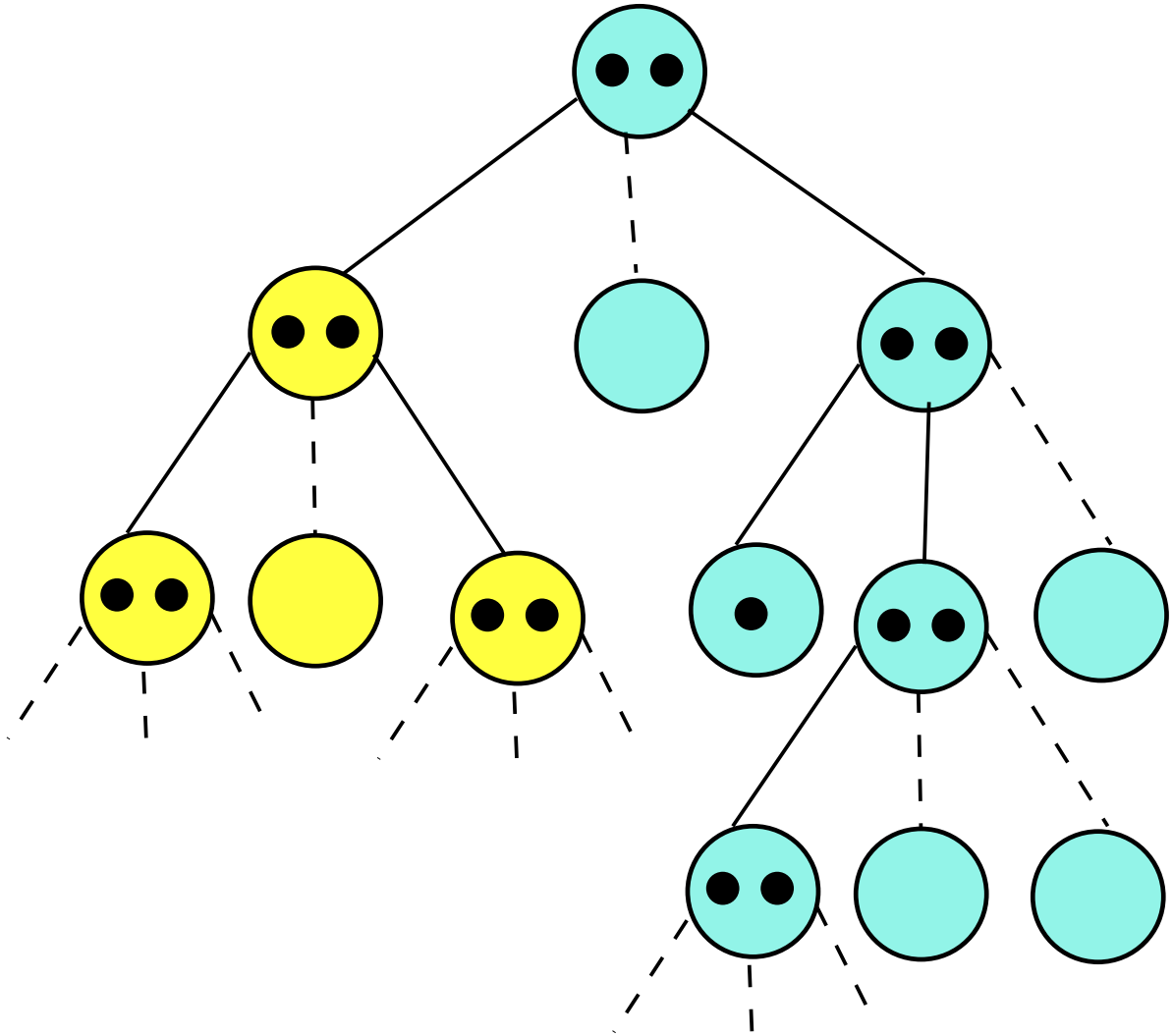
The order of the keys can be reconstructed. The "*internal*" nodes hold keys and the "*external*" nodes are empty.



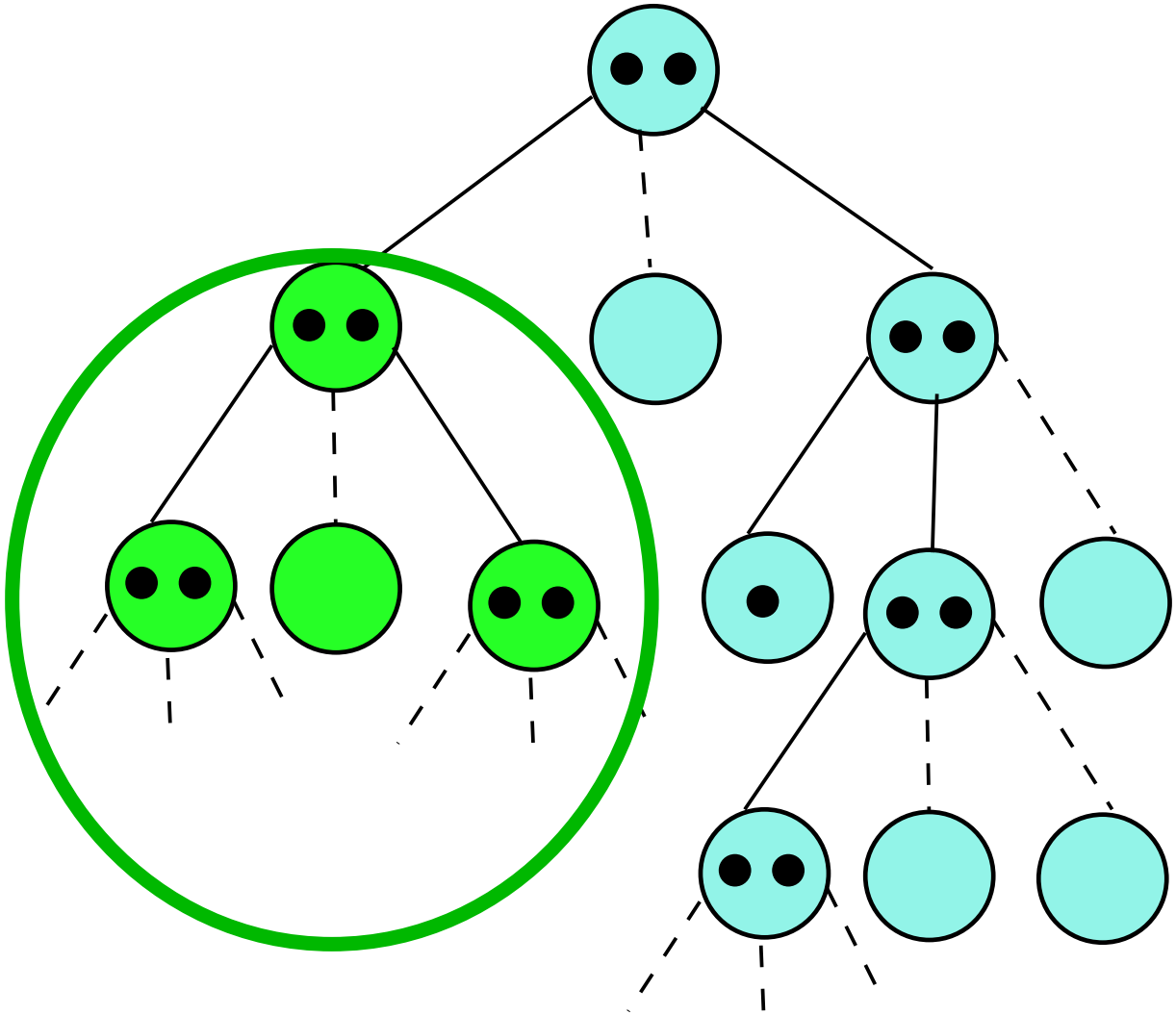
What is a fringe subtree?



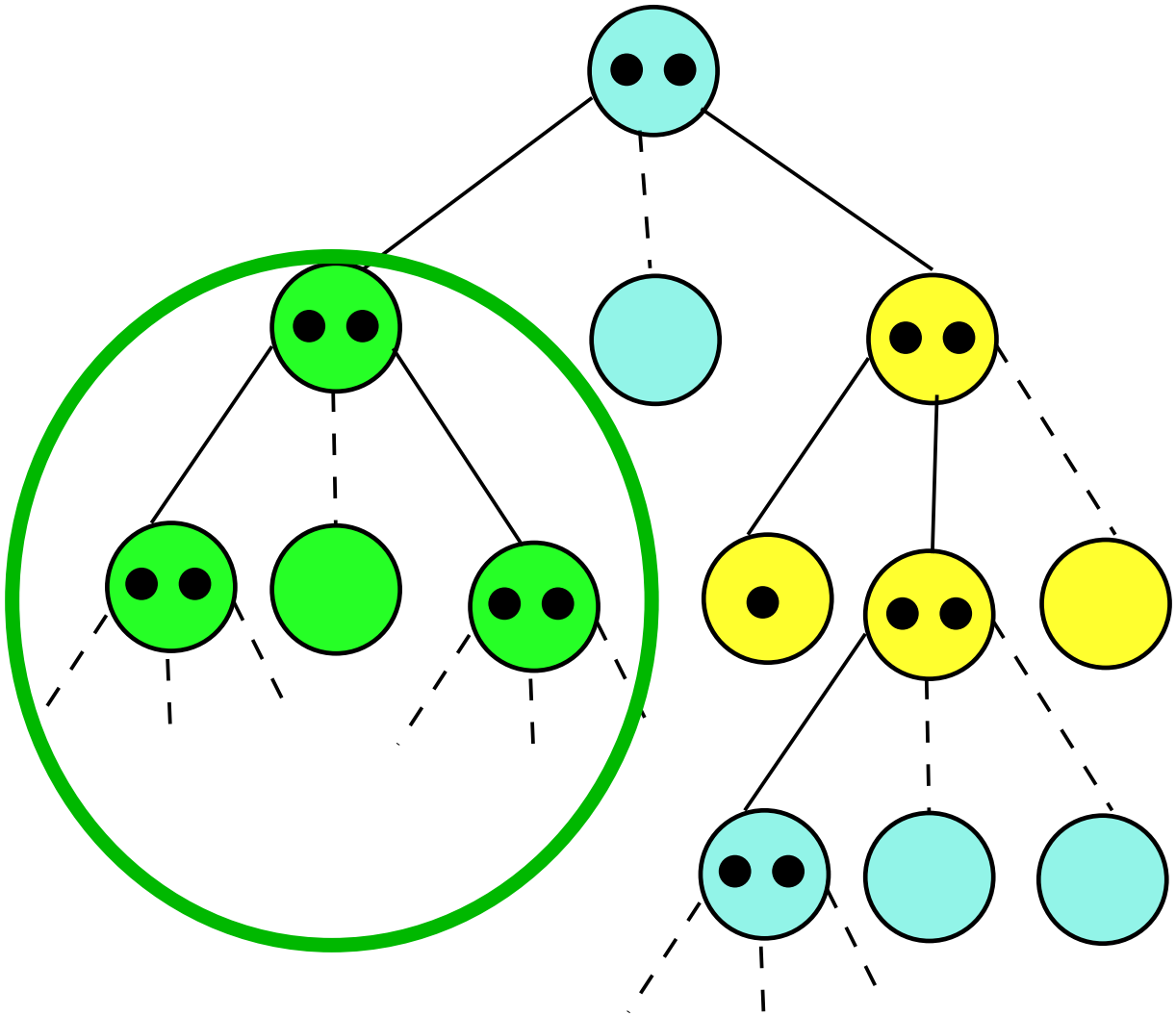
What is a fringe subtree?



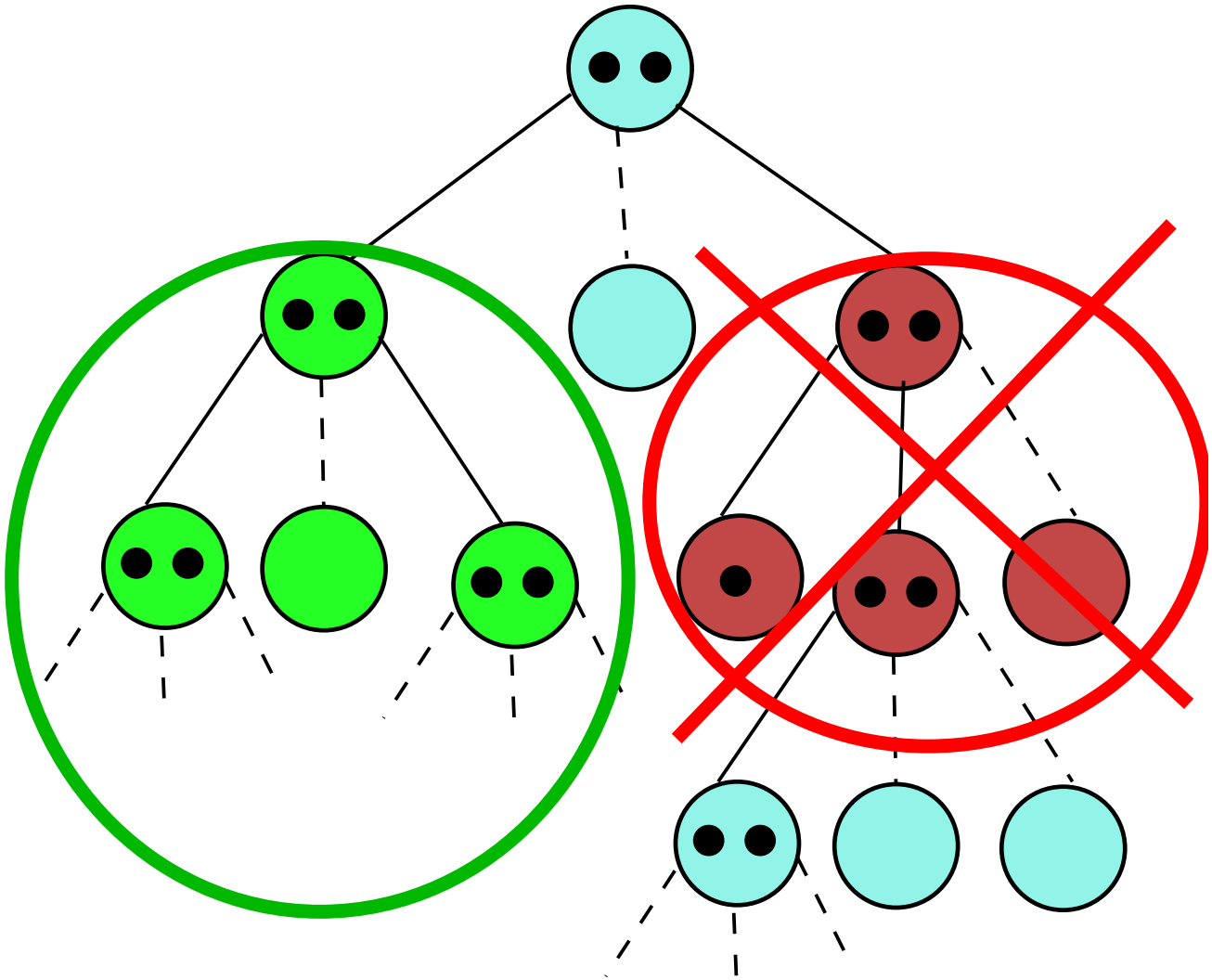
What is a fringe subtree?



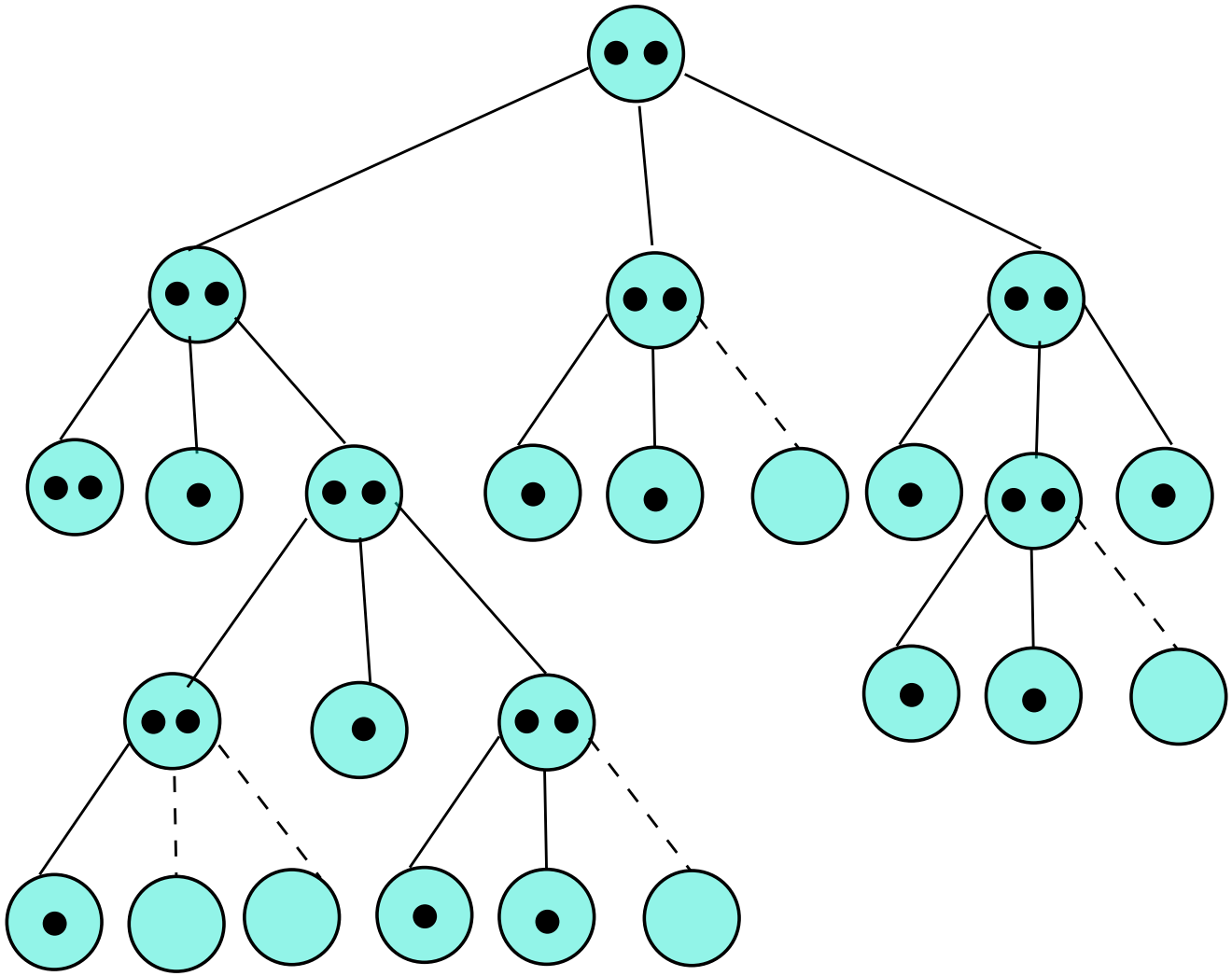
What is a fringe subtree?



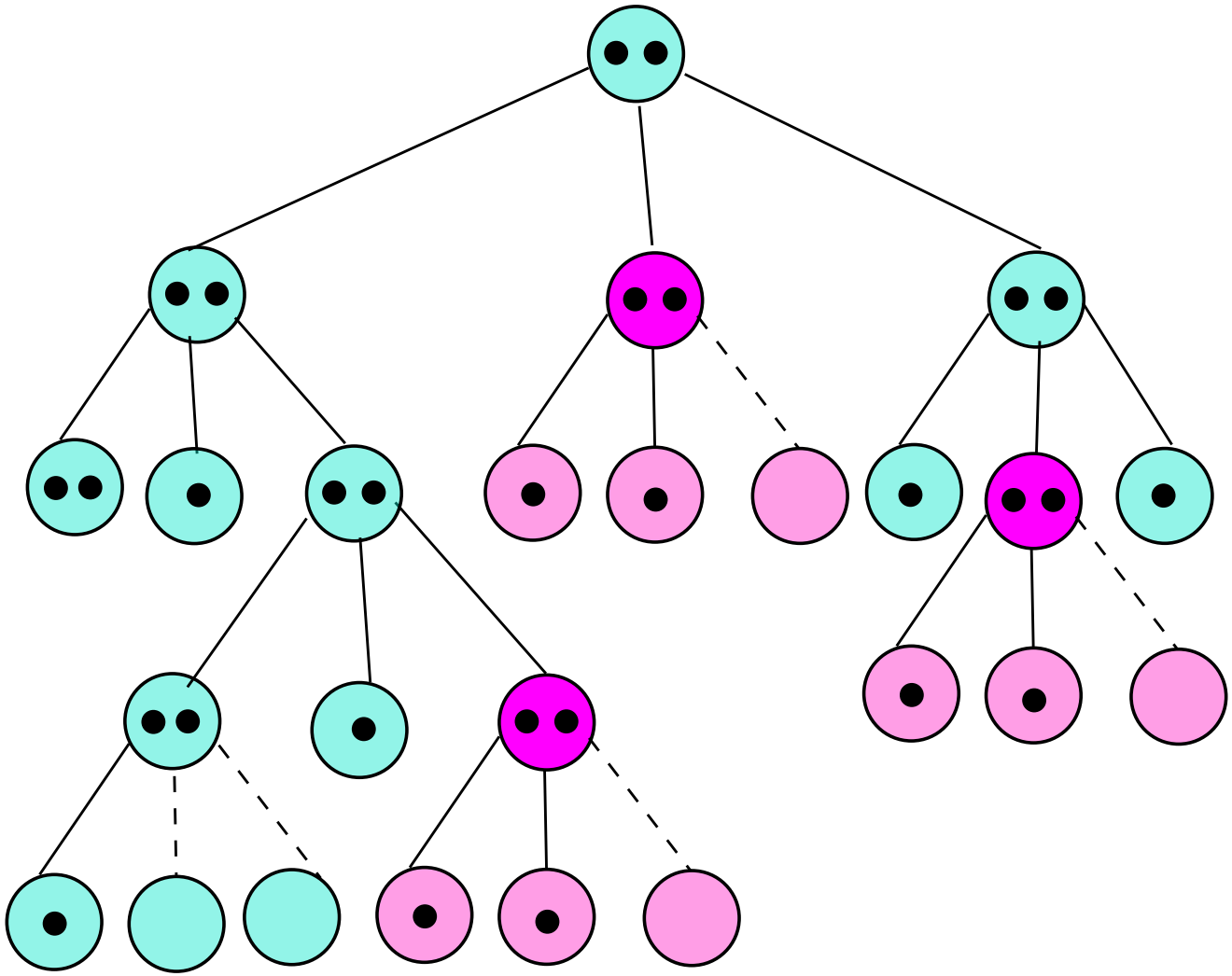
What is a fringe subtree?



Counting fringe subtrees



Counting fringe subtrees



Main Result

- Let $T^1, \dots, T^d$ be a fixed sequence of nonrandom m-ary search trees.
- Let $\mathbf{Y}_n = \left(X_n^{T^1}, X_n^{T^2}, \dots, X_n^{T^d}\right)$, where $X_n^{T^i}$ is the (random) number of fringe subtrees that are isomorphic to T^i in the random m-ary search tree $\mathcal{T}_n$ with n keys.
- Let k_i be the number of keys of T^i for $i \in \{1, \dots, d\}$.
- Let

$$\boldsymbol{\mu}_n := \mathbb{E}\,\mathbf{Y}_n = \left(\mathbb{E}(X_n^{T^1}), \mathbb{E}(X_n^{T^2}), \dots, \mathbb{E}(X_n^{T^d})\right).$$

Main Result

Recall that $\mathbf{Y}_n = \left(X_n^{T^1}, X_n^{T^2}, \ldots, X_n^{T^d}\right)$ and that $\boldsymbol{\mu}_n = \mathbb{E}\,\mathbf{Y}_n$.

Theorem

Assume that $2 \leq m \leq 26$. Then, as $n \to \infty$,

$$n^{-1/2}(\mathbf{Y}_n - \boldsymbol{\mu}_n) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma), \tag{1}$$

where $\Sigma = (\sigma_{ij})_{i,j=1}^d$ is some covariance matrix. Furthermore, in (1), the vector $\boldsymbol{\mu}_n$ can be replaced by the vector $\hat{\boldsymbol{\mu}}_n := n\hat{\boldsymbol{\mu}}$, with

$$\hat{\boldsymbol{\mu}} := \left(\frac{\mathbb{P}(\mathcal{T}_{k_1} = T^1)}{(H_m - 1)(k_1 + 1)(k_1 + 2)}, \ldots, \frac{\mathbb{P}(\mathcal{T}_{k_d} = T^d)}{(H_m - 1)(k_d + 1)(k_d + 2)}\right),$$

where $H_m := \sum_{k=1}^m 1/k$ is the m'th harmonic number. Moreover, if the trees $T^1, \ldots, T^d$ have at least one internal node each, then the covariance matrix Σ is non-singular.

Corollary of Main Result

Let k be an arbitrary fixed integer and let $Y_{n,k}$ be the (random) number of fringe subtrees with k keys in the random m-ary search tree $\mathcal{T}_n$ with n keys.

Corollary 1

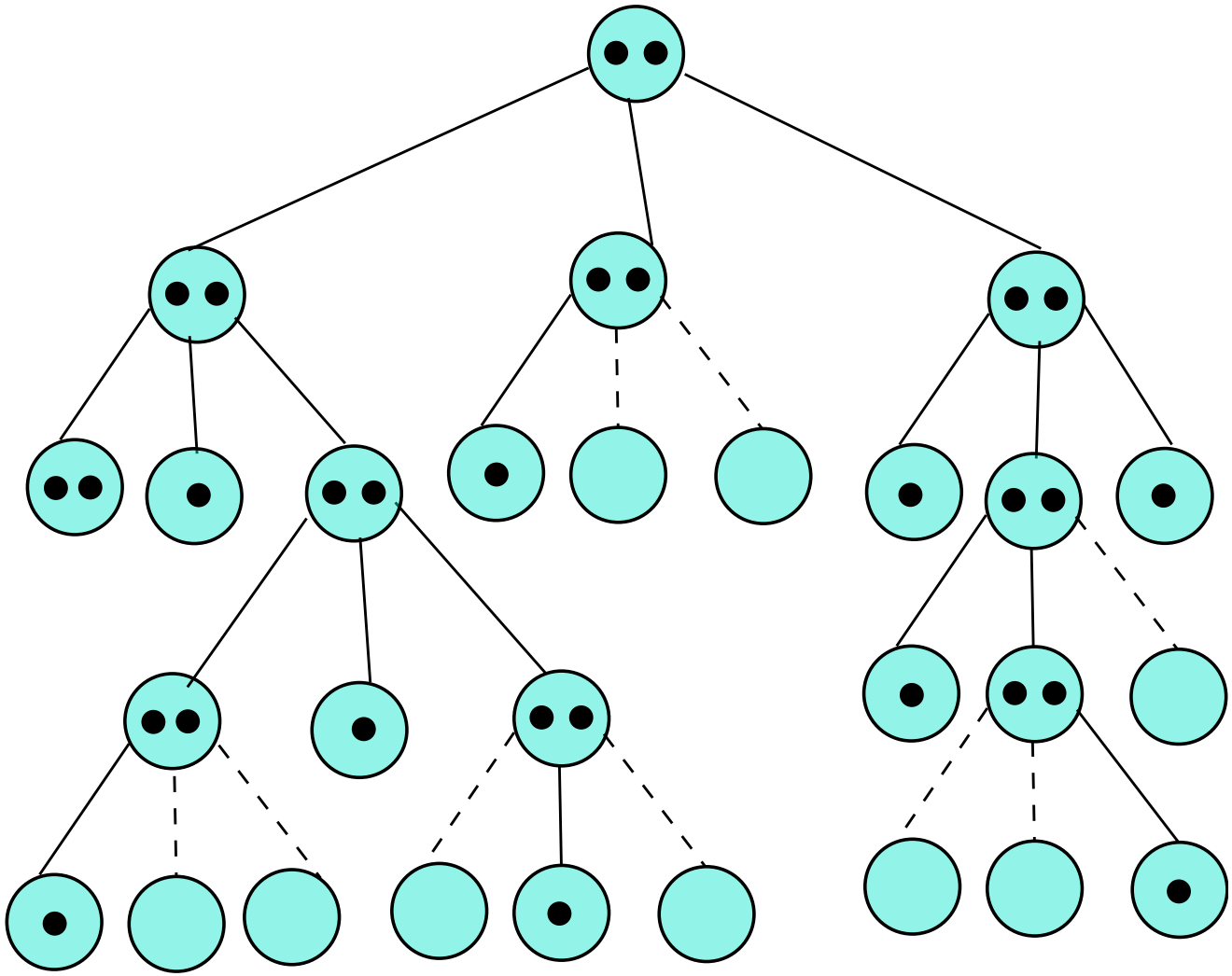
Assume that $2 \le m \le 26$. *Then, as* $n \to \infty$,

$$n^{-1/2}\bigl(Y_{n,k} - \mathbb{E}\, Y_{n,k}\bigr) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2),$$

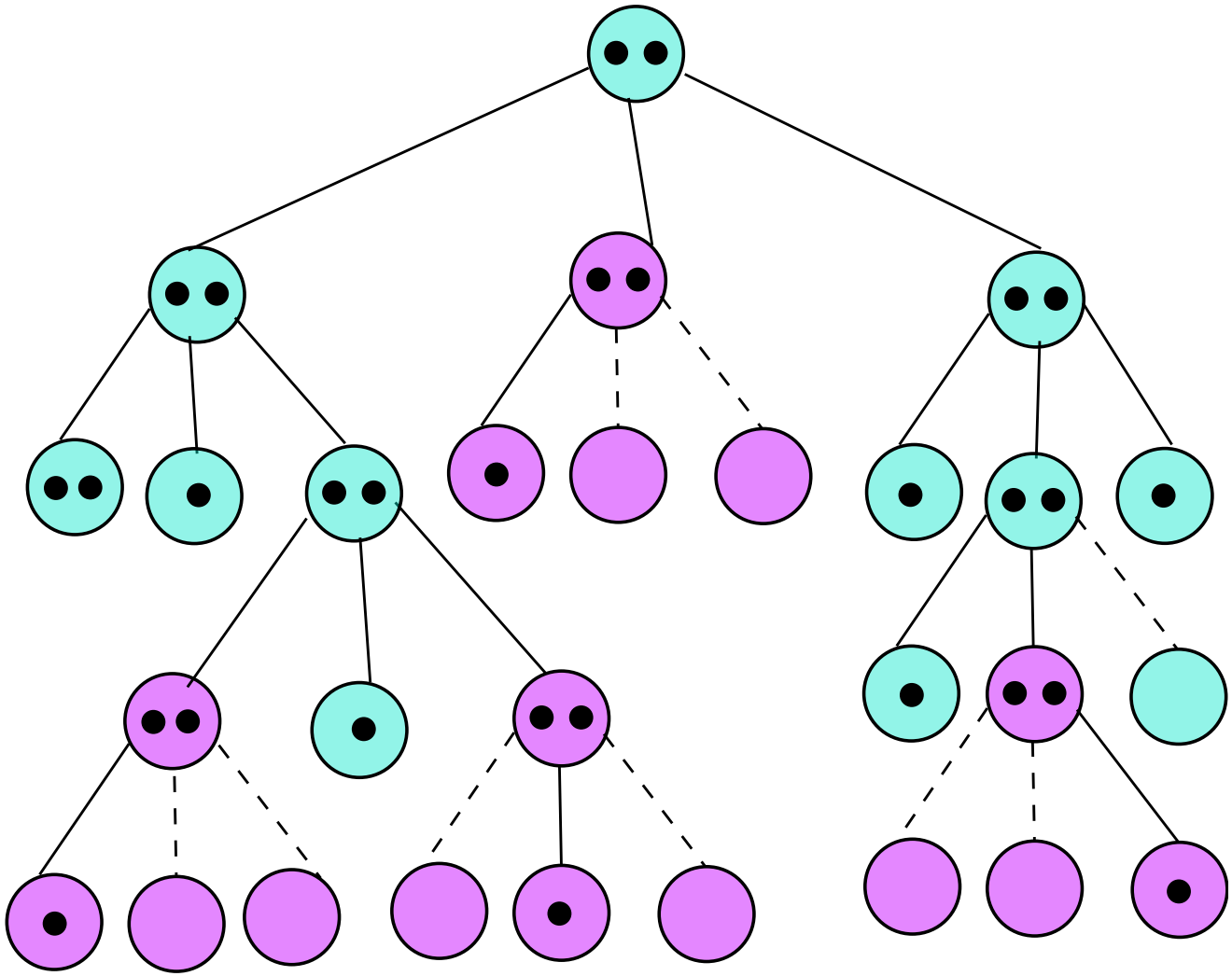
where σ_k^2 *is some constant with* $\sigma_k^2 > 0$ *except when* $k = 0$ *and* $m = 2$. *Furthermore, we also have*

$$n^{-1/2}\left(Y_{n,k} - \frac{n}{(H_m - 1)(k+1)(k+2)}\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2).$$

Counting fringe subtrees of size 3



Counting fringe subtrees of size 3



Generalised Pólya urns

- There are balls of q types (or colours) $1, \ldots, q$, and for each n a random vector $\mathcal{X}_n = (X_{n,1}, \ldots, X_{n,q})$, where $X_{n,i}$ is the number of balls of type i in the urn at time n.

- The urn starts with a given vector $\mathcal{X}_0$. Each type i is given an activity $a_i \geq 0$ and a random vector $\xi_i = (\xi_{i1}, \ldots, \xi_{iq})$, which describes the change of the composition of balls in the urn when a ball of type i is drawn. (In fact it often happens that ξ_i is deterministic, and thus the randomness in the urn process only comes from drawing of the balls.) We will assume that $\xi_{ii} \geq -1$ and $\xi_{ij} \geq 0, i \neq j$.

- The urn evolves according to a discrete time Markov process. At each time $n \geq 1$, one ball is drawn at random, with the probability of any ball proportional to its activity.

Generalised Pólya urns

- If the drawn ball has type i, it is replaced by $\Delta X_{n,j}^{(i)}$ balls of type j, $j = 1, \ldots, q$, where the random vector $\Delta X_n^{(i)} = (\Delta X_{n,1}^{(i)}, \ldots, \Delta X_{n,q}^{(i)})$ has the same distribution as $\xi_i = (\xi_{i1}, \ldots, \xi_{iq})$. (We allow $\Delta X_{n,i}^{(i)} = -1$, which means that the drawn ball is *not* replaced.)

- We let A denote the $q \times q$ matrix

$$A = (a_j \, \mathbb{E}\, \xi_{ji})_{i,j=1}^{q}.$$

The intensity matrix A with its eigenvalues and eigenvectors is central for proving limit theorems for $\mathcal{X}_n = (X_{n,1}, \ldots, X_{n,q})$.

Generalised Pólya urns and normal limit theorem

The following theorem holds under some conditions of the random process (these are often easy to verify using the Perron-Frobenius theory). Recall that $\mathcal{X}_n = (X_{n,1}, \ldots, X_{n,q})$, where $X_{n,i}$ is the number of balls of type i in the urn at time n. Let λ_1 denote the largest real eigenvalue of A and a certain right eigenvector v_1 corresponding to λ_1, i.e., $Av_1 = \lambda_1 v_1$.

Theorem (Janson (2004) Theorem 3.22)

Assume that $\mathrm{Re}\lambda < \lambda_1/2$ *for each eigenvalue* $\lambda \neq \lambda_1$.
Then, as $n \to \infty$,

$$n^{-1/2}(\mathcal{X}_n - n\mu) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

with $\mu = \lambda_1 v_1$ *and some covariance matrix* Σ.

The covariance matrix Σ is expressed in Janson (2004) and can be evaluated using the intensity matrix A.

Applying Pólya urns to count fringe subtrees

- Recall that we want to count the total number of fringe subtrees that are isomorphic to some fixed trees $T^1, \ldots, T^d$ in the random m-ary search tree $\mathcal{T}_n$.

- We will model our process as a Pólya urn, by subdividing the m-ary search tree into subtrees that represent the types $\{1, 2, \ldots, q\}$, where some of the types represent the fringe subtrees isomorphic to $T^1, \ldots, T^d$ in the tree.

- We consider adding a key to the tree in terms of the Pólya urn process of drawing a ball. In particular this will show a multivariate normal limit law for the vector $\mathbf{Y}_n = \left(X_n^{T^1}, X_n^{T^2}, \ldots, X_n^{T^d}\right)$, i.e., the number of fringe subtrees that are isomorphic to $T^1, \ldots, T^d$ in $\mathcal{T}_n$.

Types in the Pólya urn counting fringe subtrees

- Let $\mathcal{T}_n$ be a given m-ary search tree with n keys together with its external nodes (containing no keys yet). Let $\mathcal{T}_n(v)$ be the fringe subtree of $\mathcal{T}_n$ rooted at node v.

- There is a natural partial order on the set of nonrandom m-ary search trees, such that $T \preceq T'$ if T' can be obtained from T by adding keys (including the case $T' = T$).

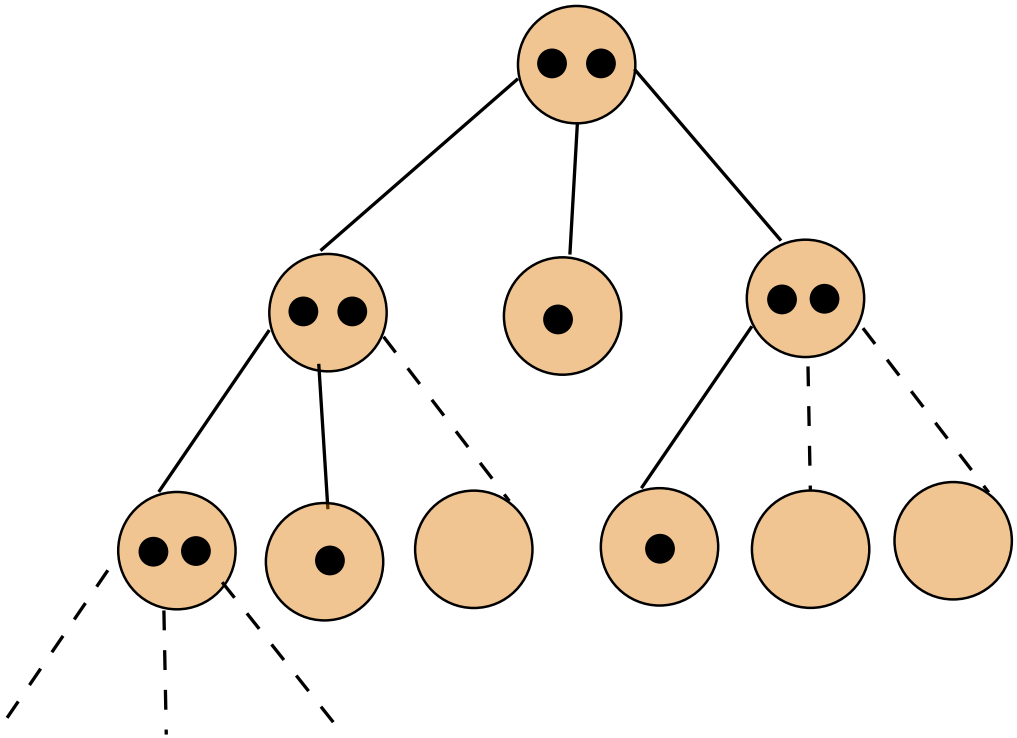
Types in the Pólya urn counting fringe subtrees

- We say that a node v is *living* if $\mathcal{T}_n(v) \preceq T^i$ for some $i \in \{1, \dots, d\}$, i.e., if $\mathcal{T}_n(v)$ is isomorphic to some T^i or can be grown to one of them by adding more keys. We let all descendants of a living node be living (all nodes of $\mathcal{T}_n(v)$ are living if v is living). All other nodes of $\mathcal{T}_n$ are *dead*.

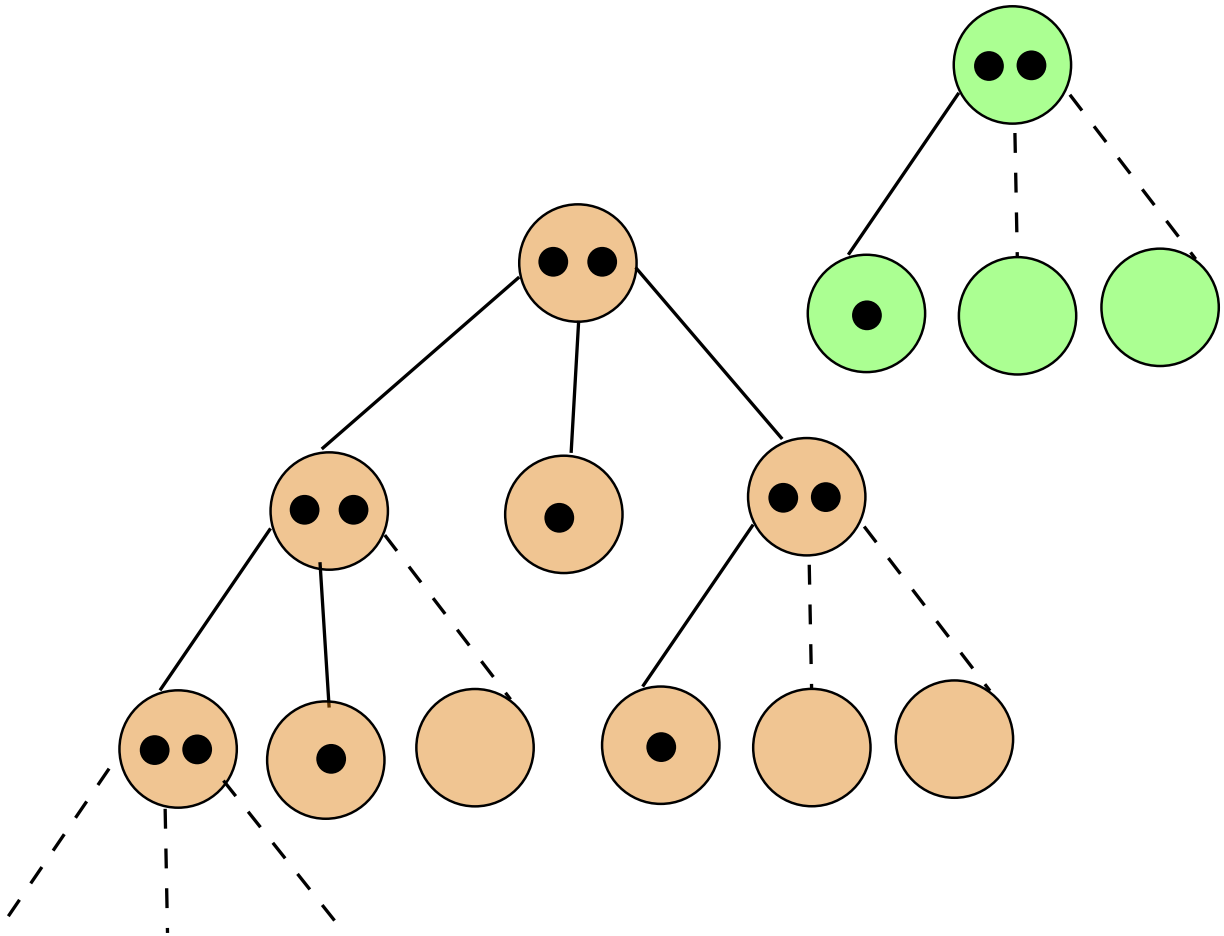
- Now erase all edges from dead nodes to their children. This yields a forest of small trees, where each tree either consists of a single dead node (with $m-1$ keys) or is living (all nodes are living) and can be grown to one of the T^i.

- We regard these small trees as the balls in the Pólya urn. Hence, the types in this urn are all nonrandom m-ary search trees T such that $T \preceq T^i$ for some $i \in \{1, \dots, d\}$, plus one dead type.

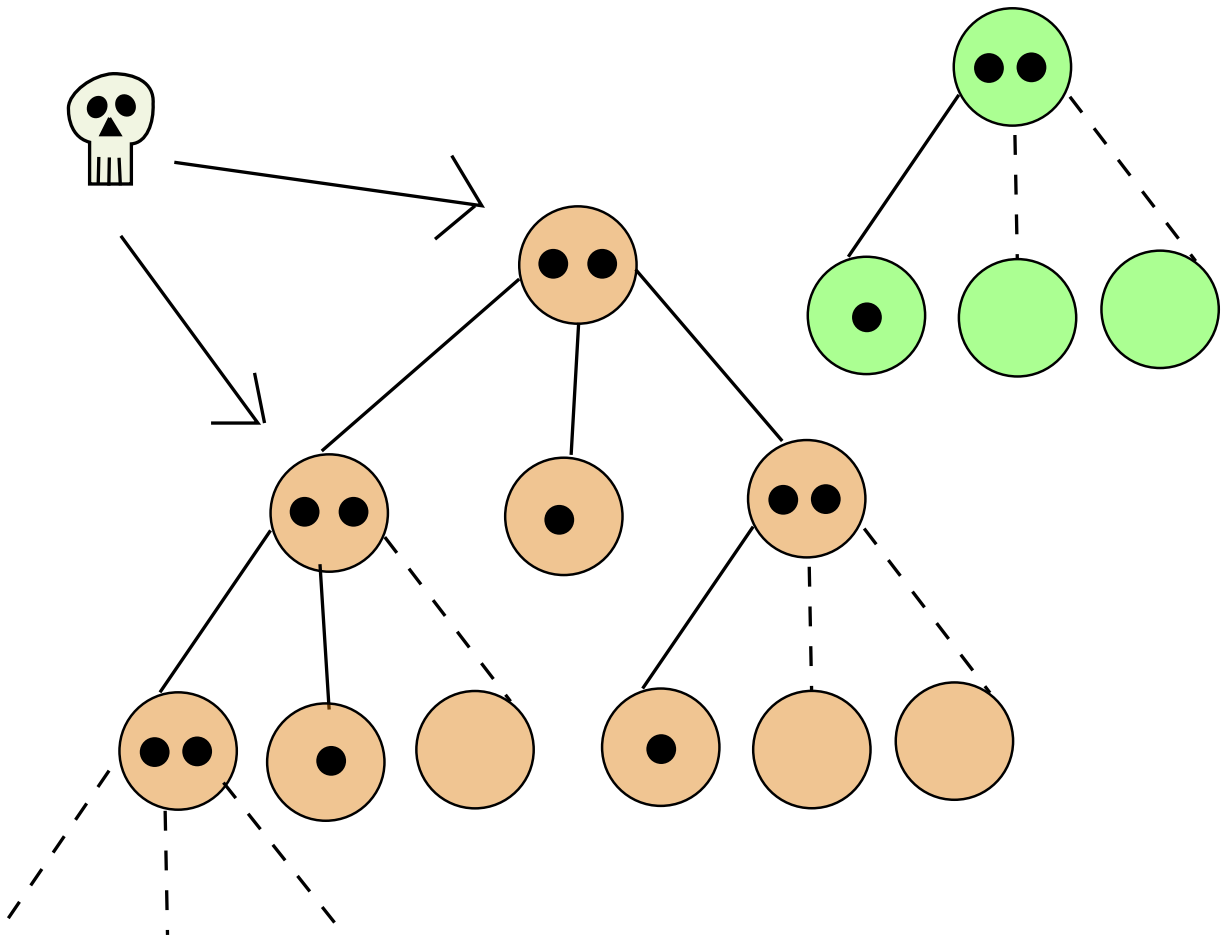
Types in the Pólya urn counting fringe subtrees



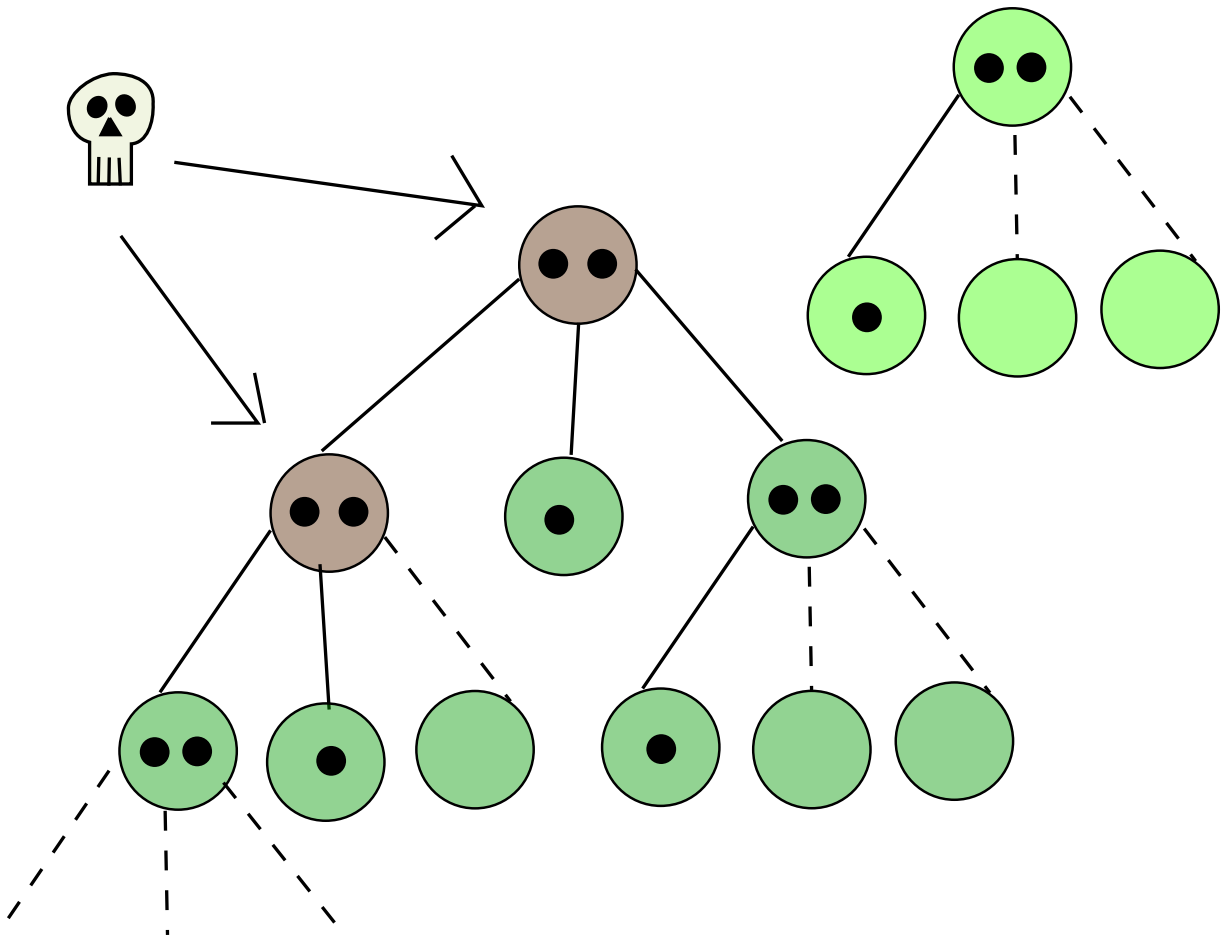
Types in the Pólya urn counting fringe subtrees



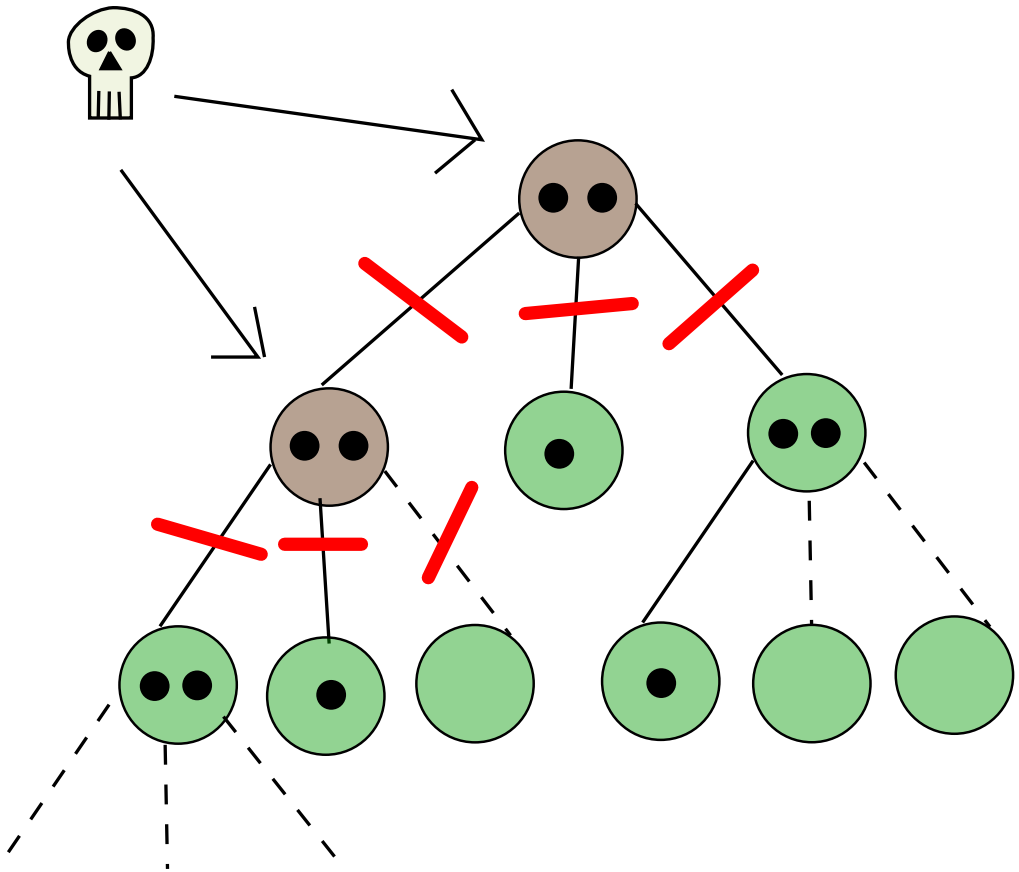
Types in the Pólya urn counting fringe subtrees



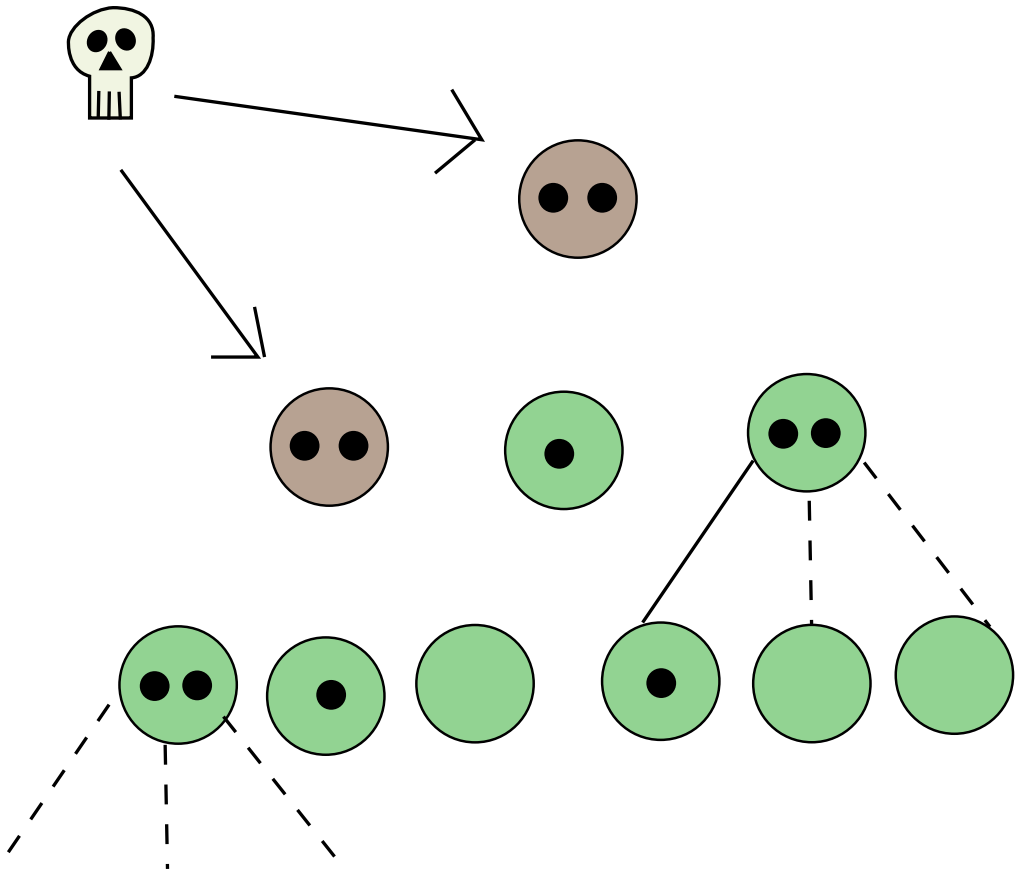
Types in the Pólya urn counting fringe subtrees



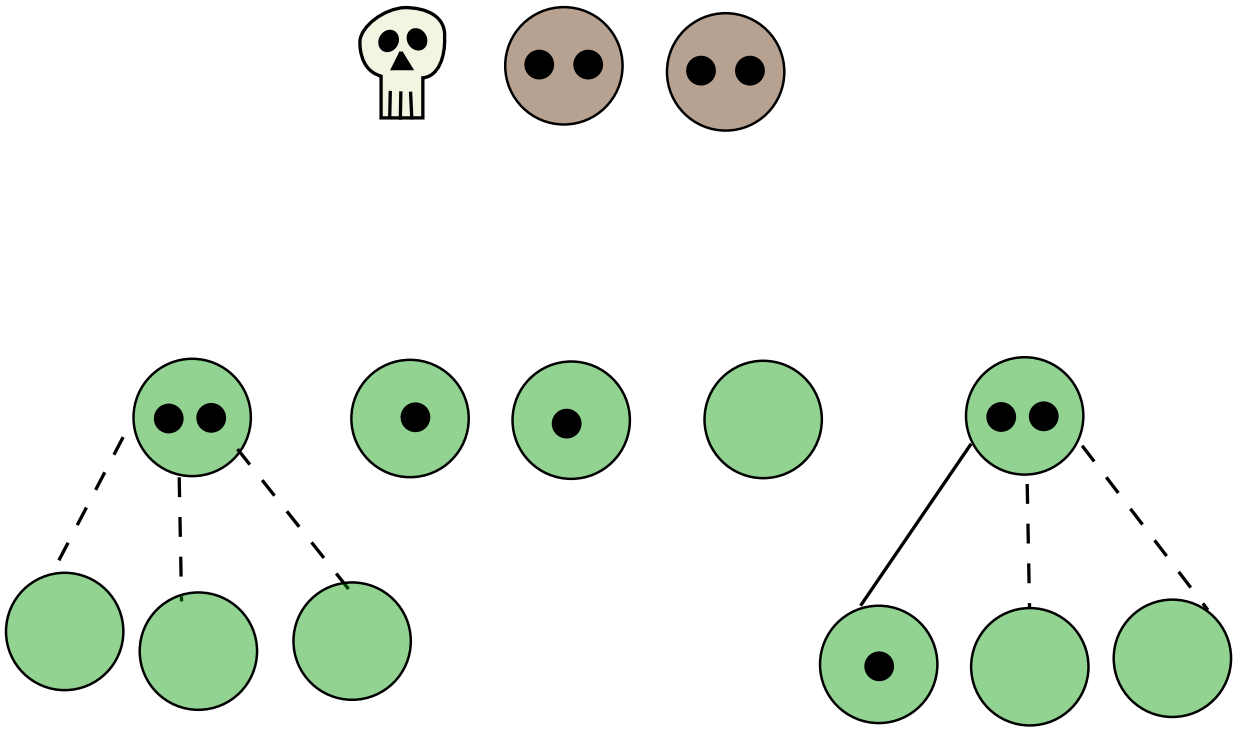
Types in the Pólya urn counting fringe subtrees



Types in the Pólya urn counting fringe subtrees



Types in the Pólya urn counting fringe subtrees



Types in the Pólya urn counting fringe subtrees

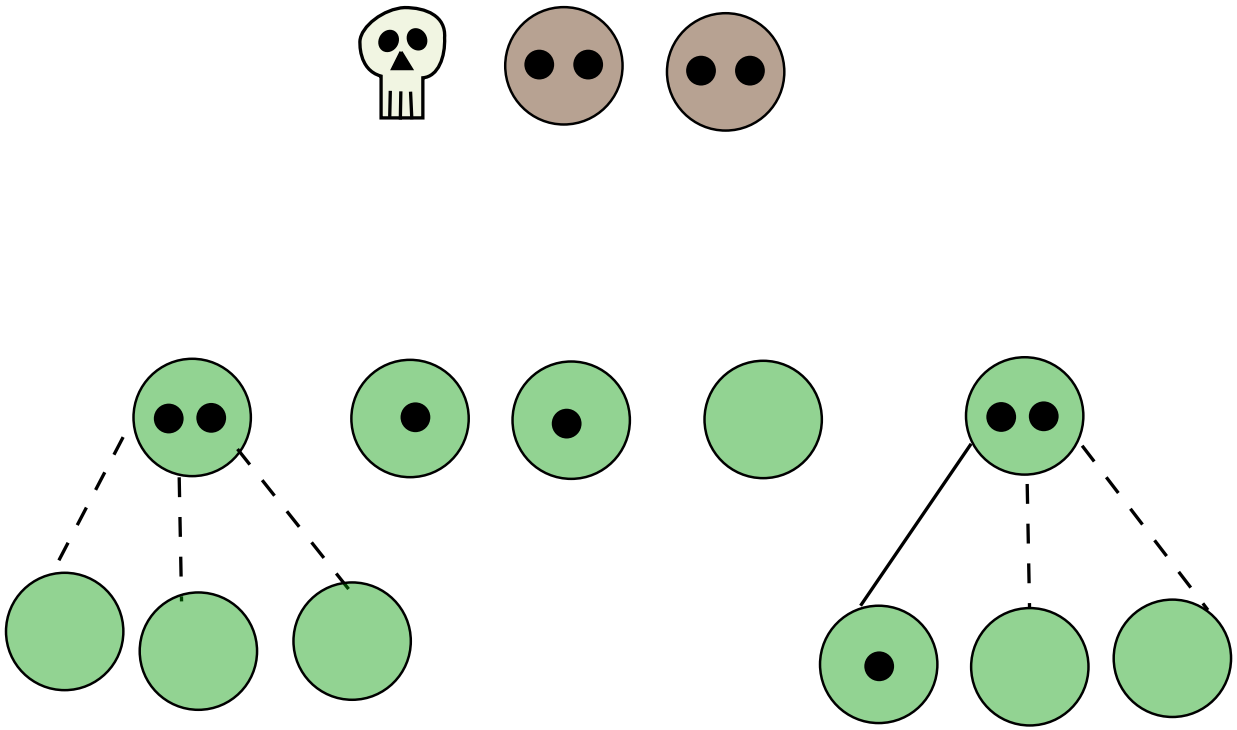
- When a key is added to the tree $\mathcal{T}_n$, it is added to a leaf with at most $m-2$ keys or an external node, and thus to one of the living subtrees.

- If the root of that subtree still is living after the addition, then that subtree becomes a living subtree of a different type; if the root becomes dead, then the subtree is further decomposed into one or several dead nodes and several (at least m) living subtrees.

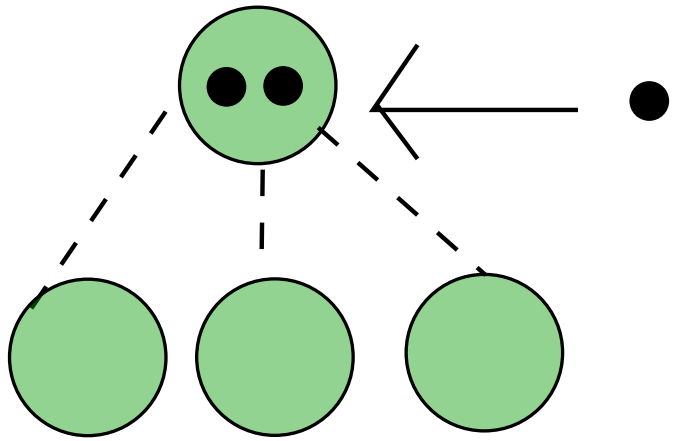
- The random evolution of the forest obtained by decomposing $\mathcal{T}_n$ is thus described by a Pólya urn, where each type has activity equal to its number of gaps.

- Note that dead balls have activity 0; hence we can ignore them and consider only the living types.

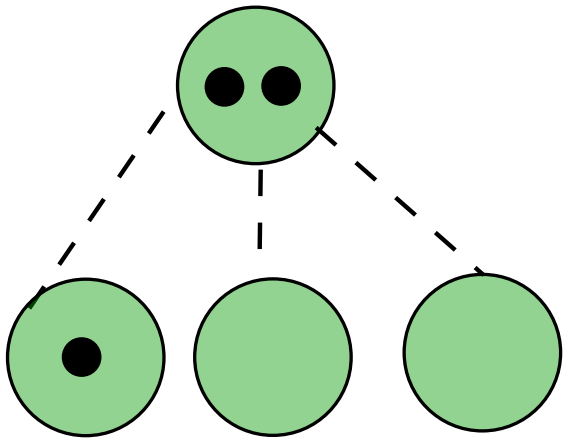
Types in the Pólya urn counting fringe subtrees



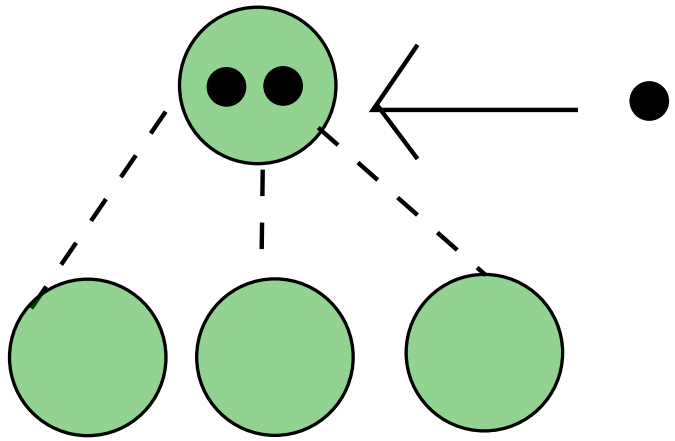
Types in the Pólya urn counting fringe subtrees



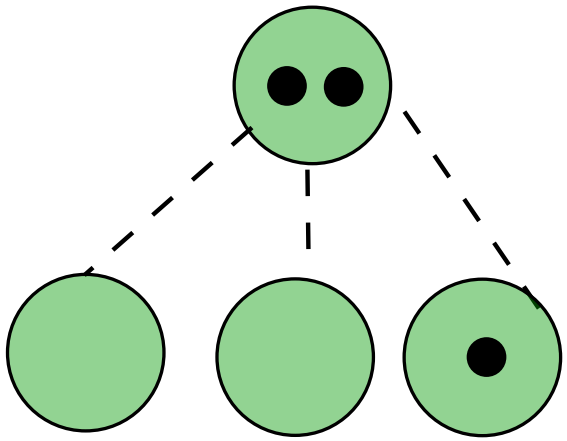
Types in the Pólya urn counting fringe subtrees



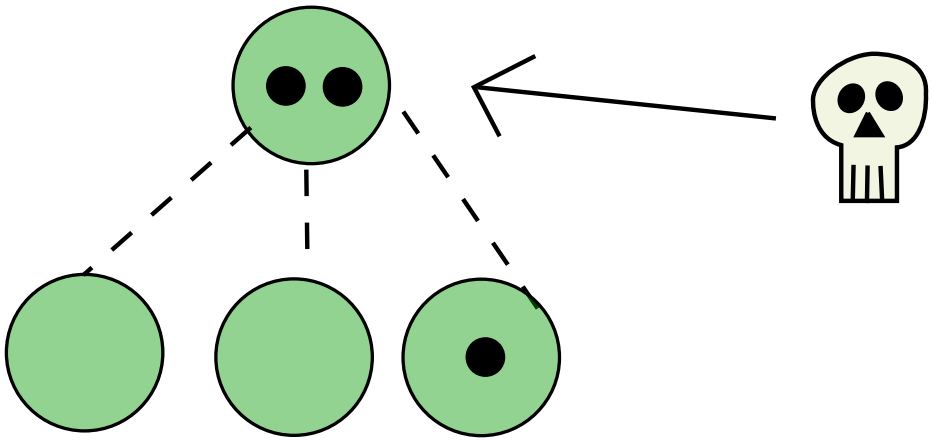
Types in the Pólya urn counting fringe subtrees



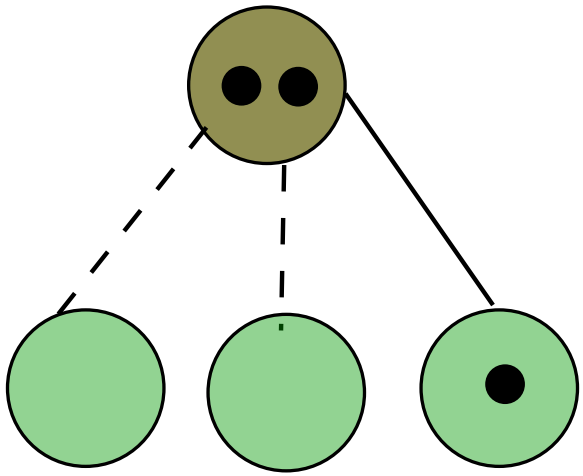
Types in the Pólya urn counting fringe subtrees



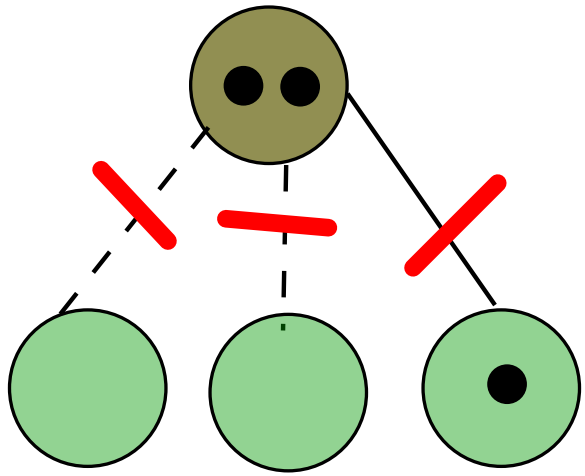
Types in the Pólya urn counting fringe subtrees



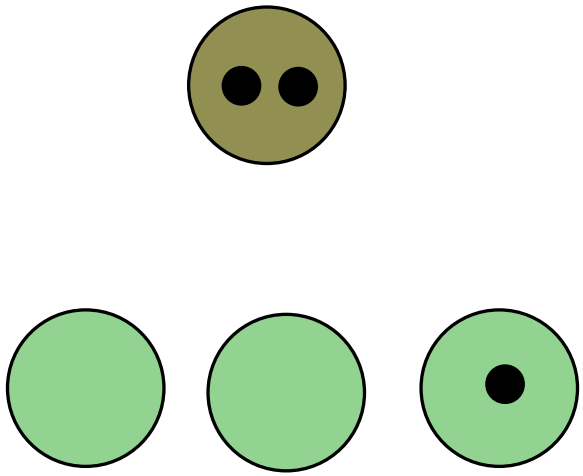
Types in the Pólya urn counting fringe subtrees



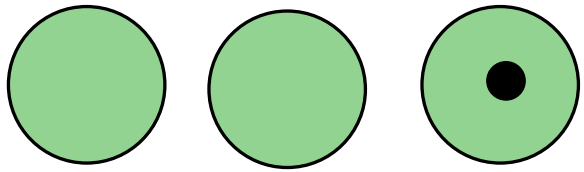
Types in the Pólya urn counting fringe subtrees



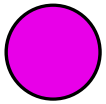
Types in the Pólya urn counting fringe subtrees



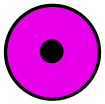
Types in the Pólya urn counting fringe subtrees



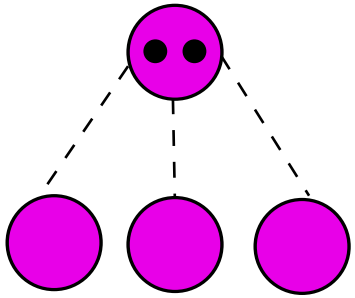
Example: Number of fringe subtrees with 4 keys



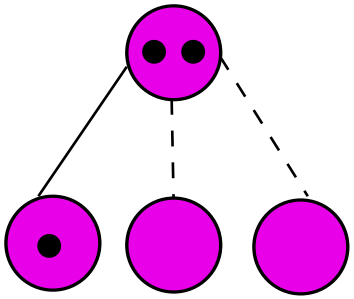
Type 1



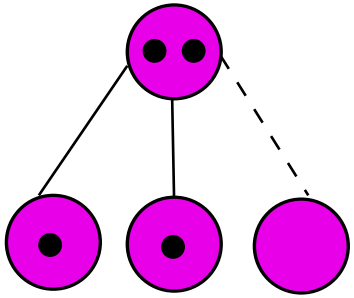
Type 2



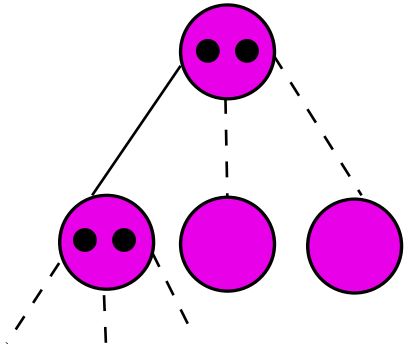
Type 3



Type 4

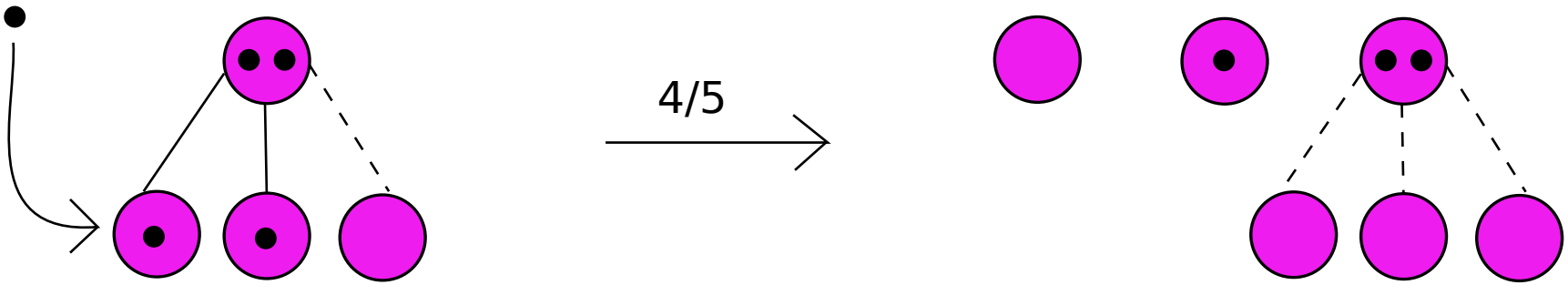


Type 5



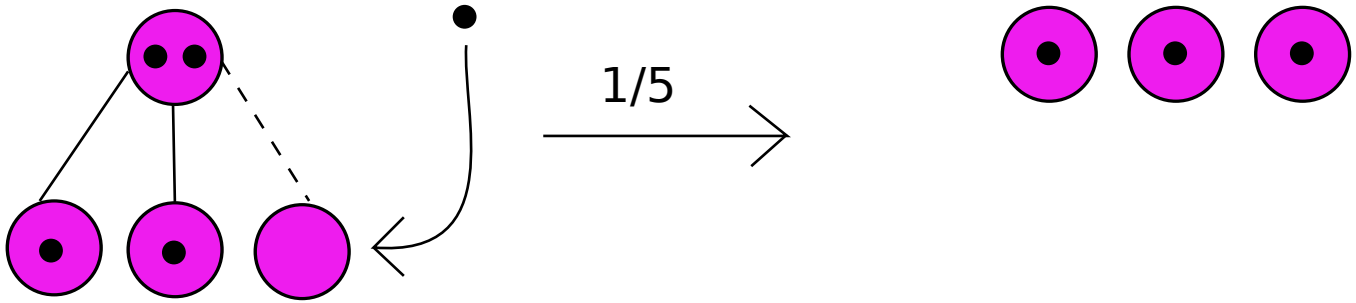
Type 6

Finding the intensity matrix A



Type 5

Type 1+Type2+Type3



Type 5

3·Type 2

Finding the intensity matrix A

Thus, we get the intensity matrix A as

$$A = \begin{pmatrix} -1 & 0 & 0 & 0 & 4 & 8 \\ 1 & -2 & 0 & 0 & 7 & 2 \\ 0 & 2 & -3 & 0 & 4 & 2 \\ 0 & 0 & 3 & -4 & 0 & 3 \\ 0 & 0 & 0 & 2 & -5 & 0 \\ 0 & 0 & 0 & 2 & 0 & -5 \end{pmatrix}.$$

The eigenvalues are $\mathbf{1}, -\mathbf{3}, -\mathbf{4}, -\mathbf{4}, -\mathbf{5}, -\mathbf{5}$.

Theorem (Janson (2004) Theorem 3.22)

Assume that $\mathrm{Re}\lambda < \lambda_1/2$ *for each eigenvalue* $\lambda \neq \lambda_1$. *Then, as* $n \to \infty$,

$$n^{-1/2}(\mathcal{X}_n - n\mu) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

with $\mu = \lambda_1 v_1$ *and some covariance matrix* Σ.

The eigenvalues for the general intensity matrices

- To show asymptotic normality in general Polya urns one needs to check $\mathrm{Re}\lambda < \lambda_1/2$ for each eigenvalue $\lambda \neq \lambda_1$ (e.g., Janson 2004).

- We will find the eigenvalues of A by using induction on the size of the set of living types.

- Note that there is exactly one type that has activity j for every $j \in \{1, \ldots, m-1\}$. (These correspond to the nodes holding $j-1$ keys.) These types are the $m-1$ smallest in the partial order $\preceq$, and they always belong to each Pólya urn counting fringe subtrees.

The eigenvalues for the general intensity matrices

- Let q be the number of types and choose a numbering $T_1, \ldots, T_q$ of these q types that is compatible with the partial order $\preceq$. For $k \leq q$, let

$$\mathcal{S}_k := \{T_1, \ldots, T_k\}.$$

- For $k \geq m - 1$, we may thus consider the Pólya urn with the k types in $\mathcal{S}_k$ constructed by chopping the whole tree $\mathcal{T}_n$ into a forest of living small trees in $\mathcal{S}_k$. Let A_k be the intensity matrix of this Pólya urn. Thus $A = A_q$.

The eigenvalues for the general intensity matrices

Proposition

Let $m \geq 3$ and $m-1 \leq k \leq q$. Then $(A_k)_{ii} = -a_i$ for every type $i = 1, \ldots, k$. Hence, the trace satisfies

$$\operatorname{tr}(A_k) = -\sum_{i=1}^{k} a_i.$$

Applying this result we show by induction on k that the eigenvalues correspond to the activities of the types in the urn.

Theorem

Let $m \geq 2$. The eigenvalues of A_q are the $m-1$ roots of the polynomial $\phi_m(\lambda) := \prod_{i=1}^{m-1}(\lambda + i) - m!$ plus the multiset

$$\{-a_i : i = m, m+1, \ldots, q\}.$$

The eigenvalues for the general intensity matrices

Theorem

Let $m \geq 2$. The eigenvalues of A_q are the $m-1$ roots of the polynomial $\phi_m(\lambda) := \prod_{i=1}^{m-1}(\lambda + i) - m!$ plus the multiset

$$\{-a_i : i = m, m+1, \ldots, q\}. \tag{2}$$

- We prove by induction on $k \geq m-1$ that the theorem holds for A_k.
- We show that A_{k+1} inherits (with multiplicities) the eigenvalues of A_k.

The eigenvalues for the general intensity matrices

- The trace of a matrix is equal to the sum of the eigenvalues; hence,

$$\operatorname{tr} A_{k+1} = \lambda_1 + \cdots + \lambda_{k+1} = \operatorname{tr} A_k + \lambda_{k+1}.$$

Recall the proposition

Proposition

Let $m \geq 3$ and $m - 1 \leq k \leq q$. Then $(A_k)_{ii} = -a_i$ for every type $i = 1, \ldots, k$. Hence, the trace satisfies $\operatorname{tr}(A_k) = -\sum_{i=1}^{k} a_i$.

- Hence, this proposition implies

$$\lambda_{k+1} = \operatorname{tr}(A_{k+1}) - \operatorname{tr}(A_k) = -a_{k+1}.$$

- Thus, by induction the theorem holds for every A_k, with $m - 1 \leq k \leq q$, and in particular for $k = q$.

Summary

- We studied fringe subtrees in random m-ary search trees, by putting them in context of generalised Pólya urns.
- We showed that for $m \leq 26$ the number of fringe subtrees that are isomorphic to $T^1, \ldots, T^d$ in $\mathcal{T}_n$ in the random m-ary search tree has a multivariate normal distribution.
- As a corollary we showed that the number of fringe subtrees with k keys in such a tree has a normal distribution.
- By using induction and the traces of the intensity matrices we saw that the eigenvalues for any Pólya urn corresponding to a set of finite fringe subtrees in the m-ary search tree correspond to the activities of the types in the urns. Thus, it followed for $m \leq 26$ that $\mathrm{Re}\lambda < \lambda_1/2$.