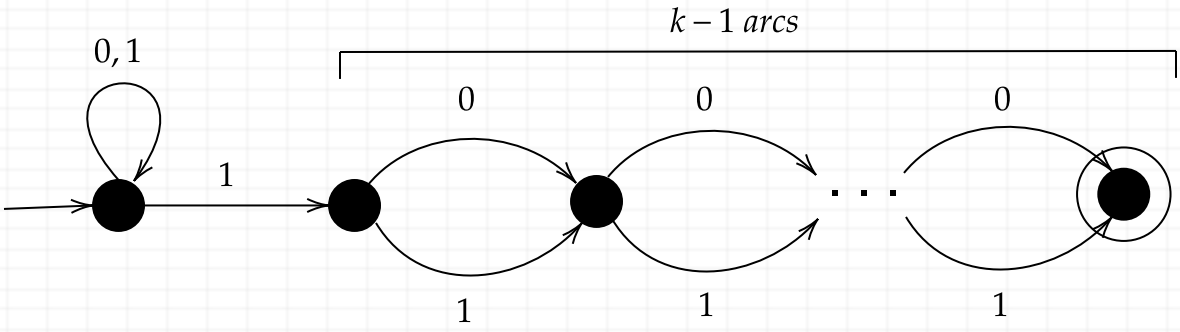
## Lecture Monday, 9/14/20: GNFAs and Regular Expressions

Another example: The "Leap of Faith" NFAs $N_k$ for any $k > 1$:




$$L(N_k) = (0+1)^* 1 (0+1)^{k-1}$$
$$= \{x \in \{0,1\}^* : \text{the kth bit of } x \text{ from the end is a } 1\}.$$

**Fact** (will be proved next week): Whereas the NFA $N_k$ has only $k+1$ states, the smallest DFA $M_k$ such that $L(M_k) = L(N_k)$ requires $2^k$ states. This is a case of **exponential blowup** in the NFA-to-DFA algorithm.

Now here is a simple algorithm for telling whether a given string $x$ *matches* a given regexp $\alpha$:

1. Convert $\alpha$ into an equivalent NFA $N_\alpha$.
2. Convert $N_\alpha$ into an equivalent DFA $M_\alpha$.
3. Run $M_\alpha$ on $x$. If it accepts, say "*yes*, it matches", else say "no match".

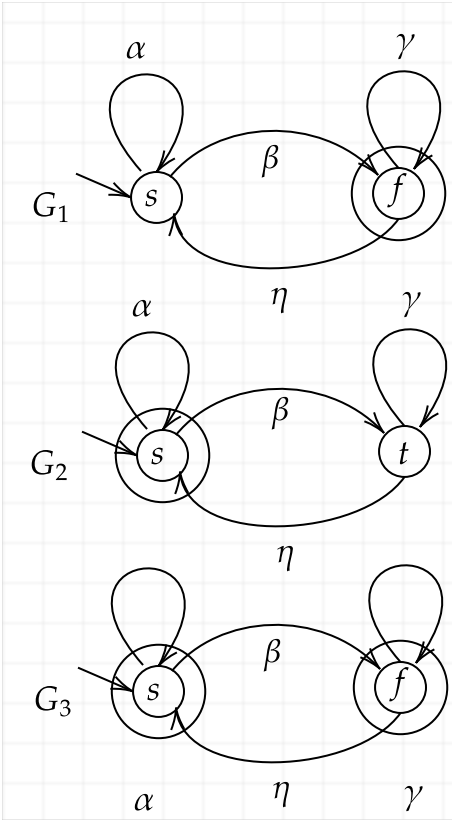
This algorithm is *correct*, but it is *not efficient*. The reason is that step 2 can blow up. An intuitive reason for the gross inefficieincy is that step 2 makes you create in advance all the "set states" that would ever be used on all possible strings $x$, but most of them are unnecessary for the particular $x$ that was given.

There is, however, a better way that builds just the set-states $R_1, \ldots, R_i, \ldots, R_n$ that are actually encountered in the particular computation on the particular $x$. We have $R_0 = S = E(s)$ to begin with. To build each $R_i$ from the previous $R_{i-1}$, iterate through every $q \in R_{i-1}$ and union together all the sets $\underline{\delta}(q, x_i)$. If $N_\alpha$ has $k$ states---which roughly equals the number of operations in $\alpha$---then that takes order $n \cdot k \cdot k$ steps. This is at worst cubic in the length $\widetilde{O}(n+k)$ of $x$ and $\alpha$ together, so this counts as a **polynomial-time algorithm**. It is in fact the algorithm actually used by the `grep` command in Linux/UNIX.

## Generalized NFAs (GNFAs) --- having only 2 states.

A *generalized NFA* $G$ can have any regular expression on any arc. A string $x$ is "accepted" by $G$ if it can be broken into $m$ substrings such that each substring matches the respective regexp in a path of $m$ arcs of $G$ that begins at $s$ and ends in a final state $f$. A regular NFA in in fact a GNFA in which every arc has a "basic"
regular expression---that is, just a char **c** in $\Sigma$, or $\epsilon$.

I do not regard GNFAs as "machines" that can be "executed"---even in the sense where we could say that the `grep` algorithm executed the NFA $N_\alpha$ on $x$. I regard them as helpful shorthand for diagramming languages. Note that the above diagram for $N_k$ makes the loop at its start state look like a GNFA arc labeled $0+1$. The only reasoning for not using single arcs labeled $0+1$ or $0,1$ in the iterated part of the machine is to emphasize the contrast with the single arc out of the start state labeled $1$ only. With concatenation and star the shorthand is more substantial. The most illuminating case IMHO of this is for two-state GNFAs:




$$L(G_1) = L_{sf} = (\alpha + \beta\gamma^*\eta)^*\beta\gamma^*$$
$$= \alpha^*\beta(\gamma + \eta\alpha^*\beta)^*$$

$\eta$ is pronounced "ate-a" in the US, "eat-a" in the UK.

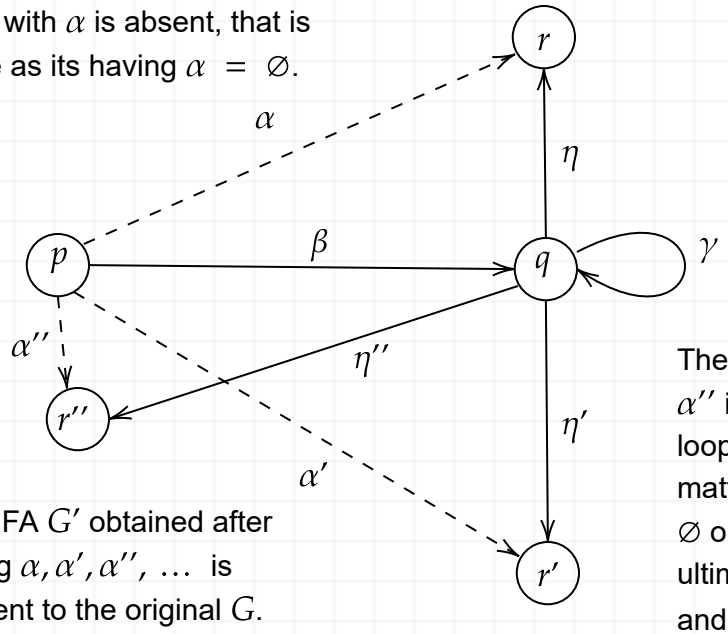
$$L(G_2) = L_{ss} = (\alpha + \beta\gamma^*\eta)^*$$

$$L(G_3) = L_{ss} \cup L_{sf}$$

Note that in $G_2$, we called the second state $t$ rather than $f$ because it is not accepting. No accepting computation can begin or end at a non-final state $q$ that is different from the start state. Hence, if the computation enters $q$ from some state $p$, then it must exit at some state $r$ (which can be the same as $p$). Considering multiple such states $r, r', r''$ gives us the following diagram:

## General GNFA Case:

If the arc with $\alpha$ is absent, that is the same as its having $\alpha = \varnothing$.




$$\alpha_{new} = \alpha_{old} + \beta\gamma^*\eta$$

$$\alpha'_{new} = \alpha'_{old} + \beta\gamma^*\eta'$$

$$\alpha''_{new} = \alpha''_{old} + \beta\gamma^*\eta''$$

The last works if $p = r''$ when $\alpha''$ is a self-loop at $p$. If the self-loop is absent, it turns out not to matter whether you take it to give $\varnothing$ or $\epsilon$. The reason is that it will ultimately be inside a Kleene star, and $(\varnothing + \zeta)^* = (\epsilon + \zeta)^* = \zeta^*$ for any regular expression $\zeta$ (zeta).

The GNFA $G'$ obtained after updating $\alpha, \alpha', \alpha'', \ldots$ is equivalent to the original $G$.

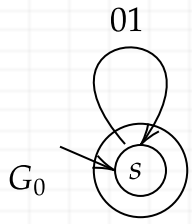
Once we have *bypassed* every edge into $q$, we can *delete* $q$.

If we are programming this with a `RegExp` package, then we can represent a given $n$-state finite automaton (DFA, NFA, or GNFA, all the same to start with) by an $n \times n$ matrix $T$ of `RegExp`. We can number the non-accepting states different from the start state by $m, \ldots, n$ for whatever $m$ applies. (If start is the only accepting state then we could take $m$ as low as 2, but it saves "mess" to take $m = 3$ in this case too so that execution will end with $G_2$ above, at which point the answer can be shortcutted by saying what $\alpha, \beta, \gamma, \eta$ are and citing the abstract formula. Most sources say to add a new start state and make all original final states go to a new one, but while doing this makes the proof look neater, it is more work that is highly typo-prone.) Then let one loop variable $k$ run over the nodes $q$ to be eliminated, let $i$ run over all states up to $k-1$ which are treated as possible entry states $p$, and let $j$ run over potential exist states $r$. Then the main code is simply:

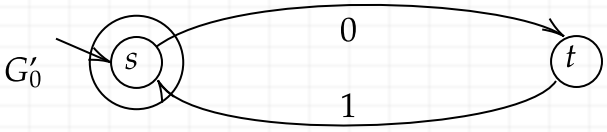
```
for k = n downto m:
    for i = 1 to k-1:
        for j = 1 to k-1:
            T(i,j) += T(i,k) · T(k,k)* · T(k,j).
```

(The convenience of writing "+=" here is one reason I like using + rather than $\cup$ for union.) Note that even if there is no self-loop at $q$, so that $T(k,k) = \varnothing$ (or $\epsilon$; it doesn't matter), the update is not killed because $T(k,k)^* = \epsilon$. But if there is no arc from $i$ into $k$, that is, if $T(i,k) = \varnothing$, then the right-hand side does get nulled and the update is simply a no-op. Likewise if no arc from $k$ out to $j$, whereupon $T(k,j) = \varnothing$.

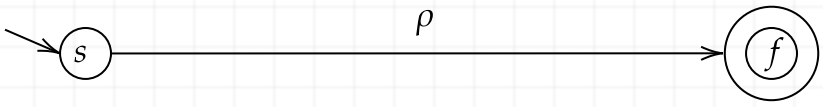
The result of executing the code is a GNFA $G'$ with all states accepting except possibly the start state. If the start state, too, is accepting, it is tempting to think $L(G') = \Sigma$, i.e., that $G'$ accepts all strings, but that is not true because GNFA arcs can have "holes" that prevent matching and hence processing all strings. For example, consider the simple one-state GNFA




Then $L(G_0) = (01)^*$ but this is not all strings. The reason is that $G_0$ was really abbreviating the NFA shown below, which can "crash" on $1$ at its start state and on $0$ at state $t$:




So if you get a $G'$ with two or more accepting states different from the start state, then you do have to add a new final state $f$ with arcs from all the old final states, declare $f$ to be the only final state, and eliminate all of the previous accepting states apart from $s$. If you also make $s$ a new, non-accepting state, then you do get the final answer $\rho = L(G)$ "on a silver platter":



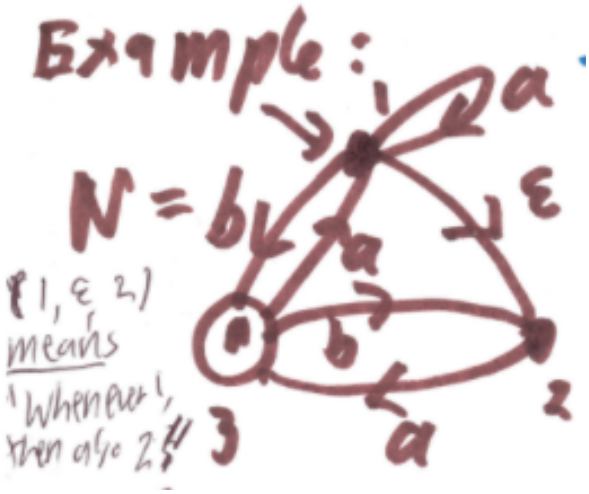

But the final expression $\rho$ you get is often quite long, and the steps for the last one or two states you eliminated often amount to hand-copying long subexpressions corresponding to $\alpha, \beta, \gamma, \eta$ in the above formulas for the 2-state GNFAs anyway. The ground rules are hence that once you get down to two states, you can just cite the abstract formula to say what the final regular expression will be. And if the originally given GNFA has at most one accpeting state besides the start state, then the above code body will give your final answer without needing to add a new final state. Why add one or two iterations to the outside of a triply-nested loop if you can avoid it?

Anyway, what we have proved is:

**Theorem.** Given any DFA, NFA, or GNFA $G$, we can calculate a regular expression $\rho$ (Greek rho) such that
$L(\rho) = L(G)$.

This also completes the proof of the final part of Kleene's Theorem.

Example---revisiting a previous NFA:

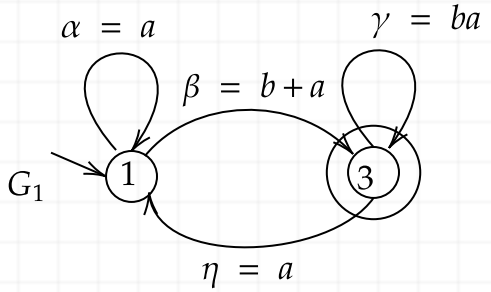



We want to eliminate state 2. If we were using the code approach, we could re-number it as state 3. But we can also do it "graphically": list the "In"coming and "Out"going arcs and update all combinations of them. Here we have:

In: 1 (on $\epsilon$) and 3 (on $b$).
Out: only to 3 (on $a$).
Update: $T(1,3)$ and $T(3,3)$.

$$\begin{aligned} T(1,3)_{new} &= T(1,3)_{old} + T(1,2)T(2,2)^*T(2,3) \\ &= b + \epsilon\cdot\epsilon\cdot a = b + a. \\ T(3,3)_{new} &= T(3,3)_{old} + T(3,2)T(2,2)^*T(2,3) \\ &= \varnothing + b\cdot\epsilon\cdot a = ba. \end{aligned}$$

The new GNFA is




$$L(G_1) = L_{sf} = \left(\alpha + \beta\gamma^*\eta\right)^*\beta\gamma^*$$

$$= (a + \left((b+a)(ba)^*a\right)^*(b+a)(ba)^*.$$

[If time permits, some philosophical discussion will follow.]