R Lab 2

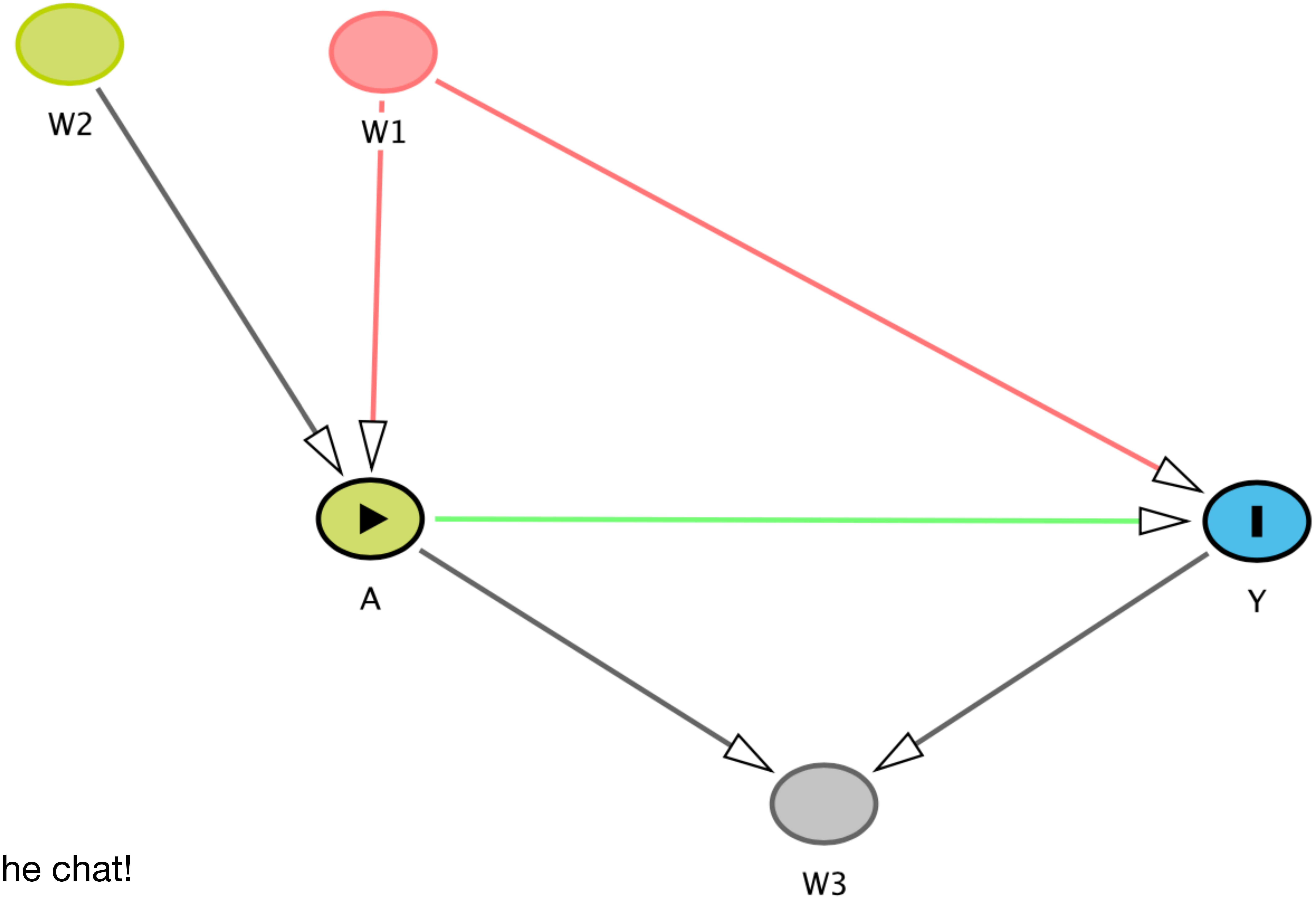
topics:

- back-door criterion
- positivity assumption
- g-computation

IMPT REMINDERS

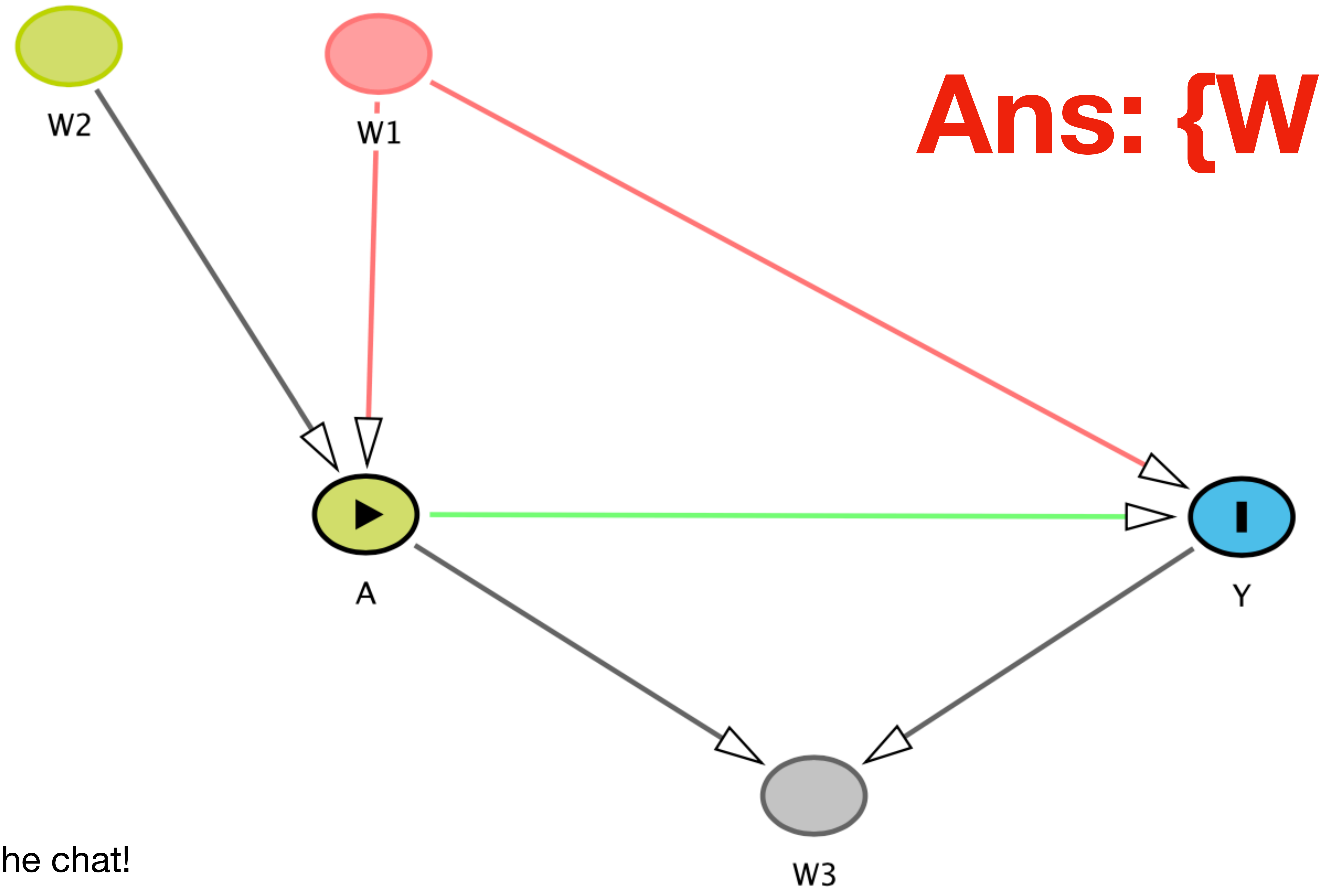
- Josh: remember to record this
- R HW 2 is posted already
- Note that TEMPLATE exists for R HW 2 so that they won't forget any parts of any questions
- Solutions to this lab will be posted soon too; R code will be very helpful

What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



Type your answer in the chat!

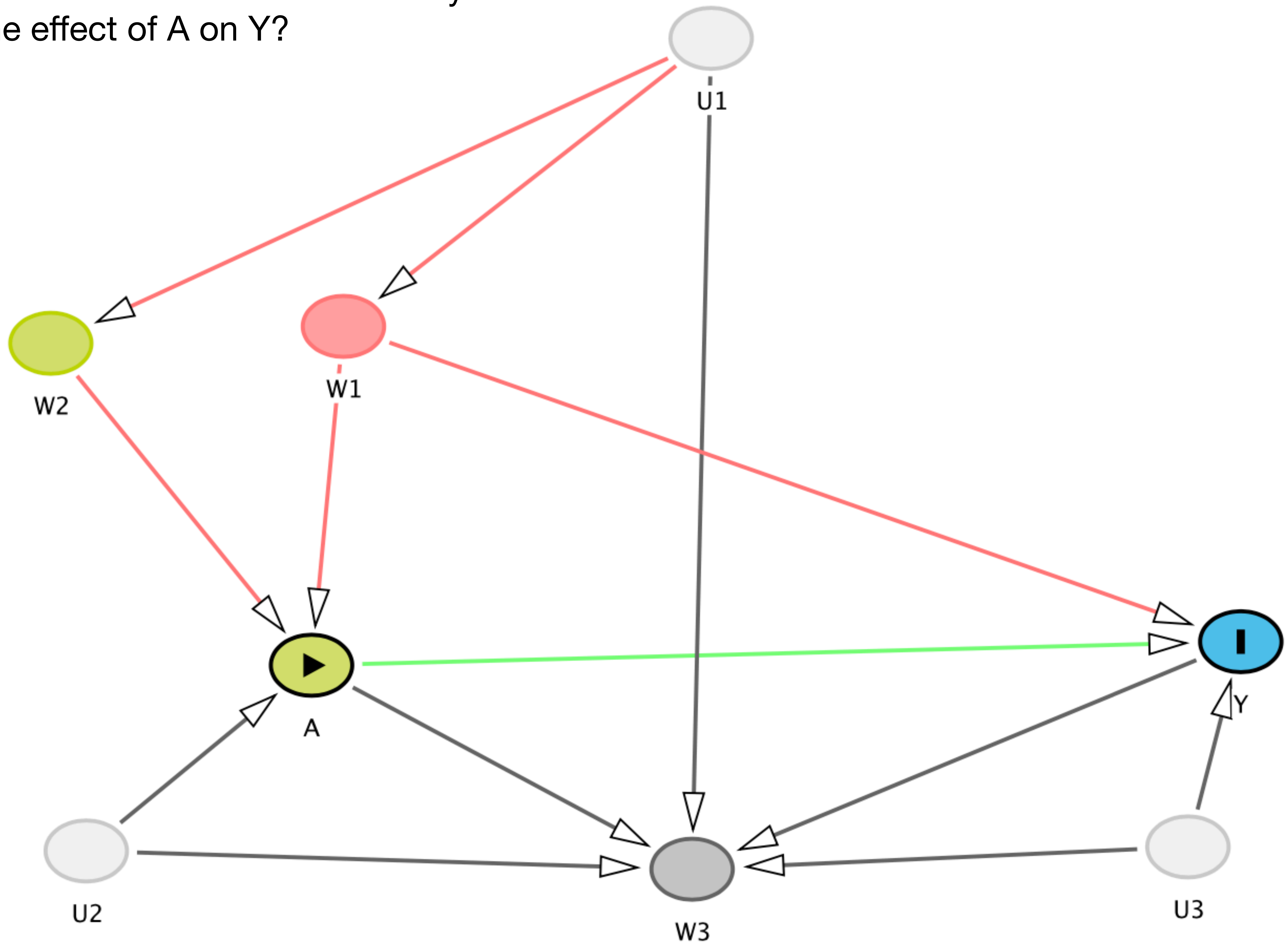
What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



Ans: {W1}

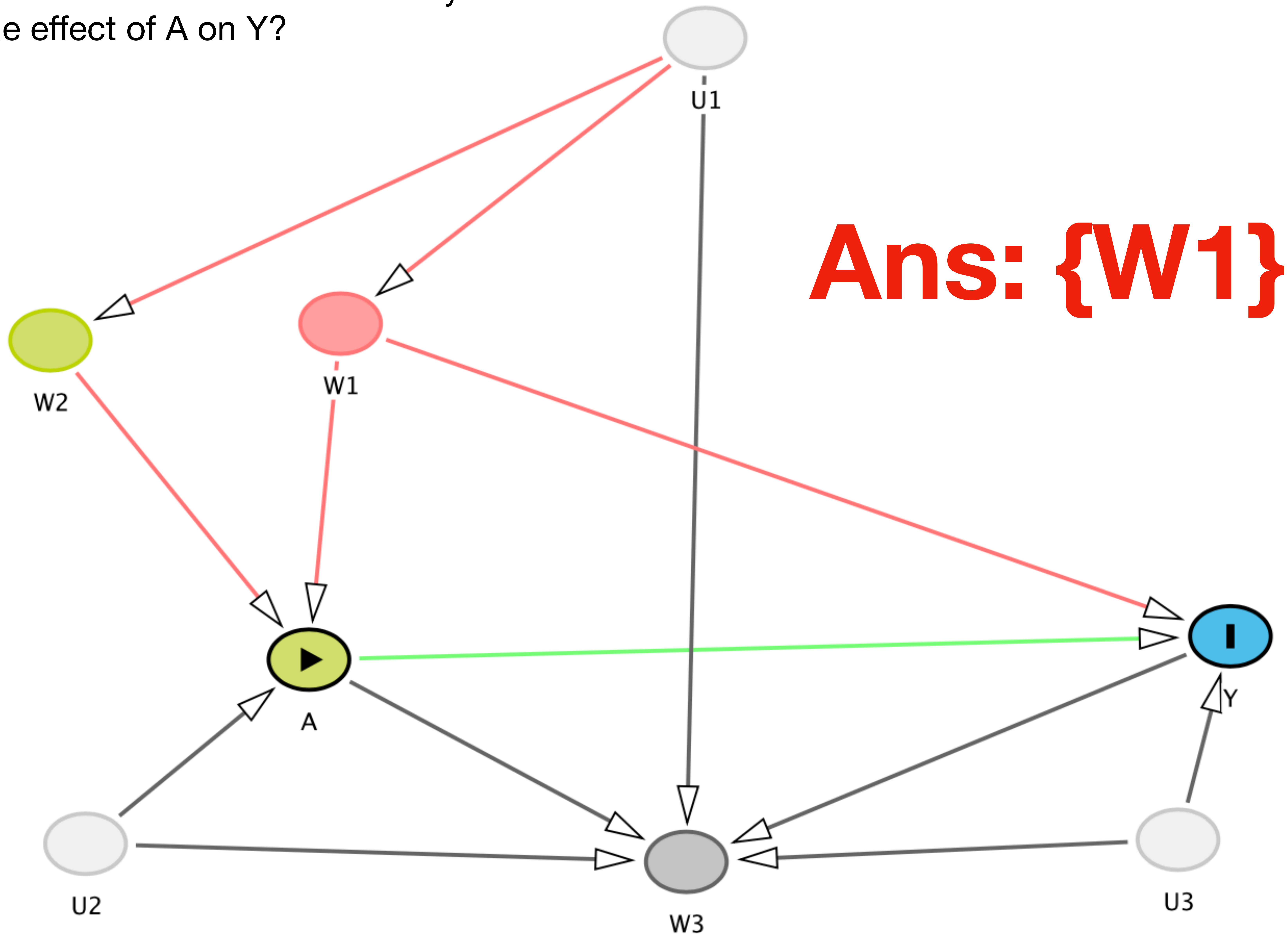
Type your answer in the chat!

What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



Type your answer in the chat!

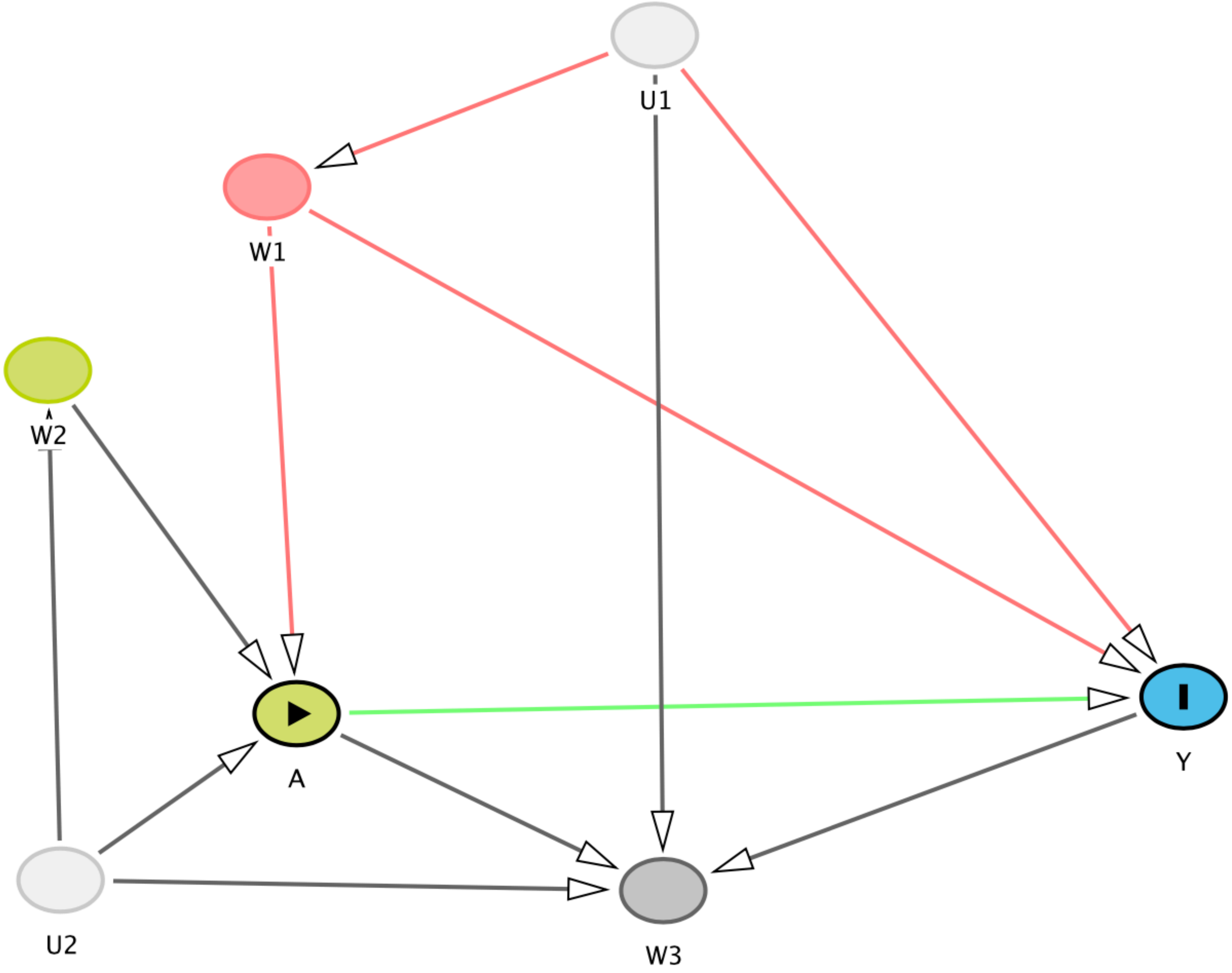
What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



Ans: {W1}

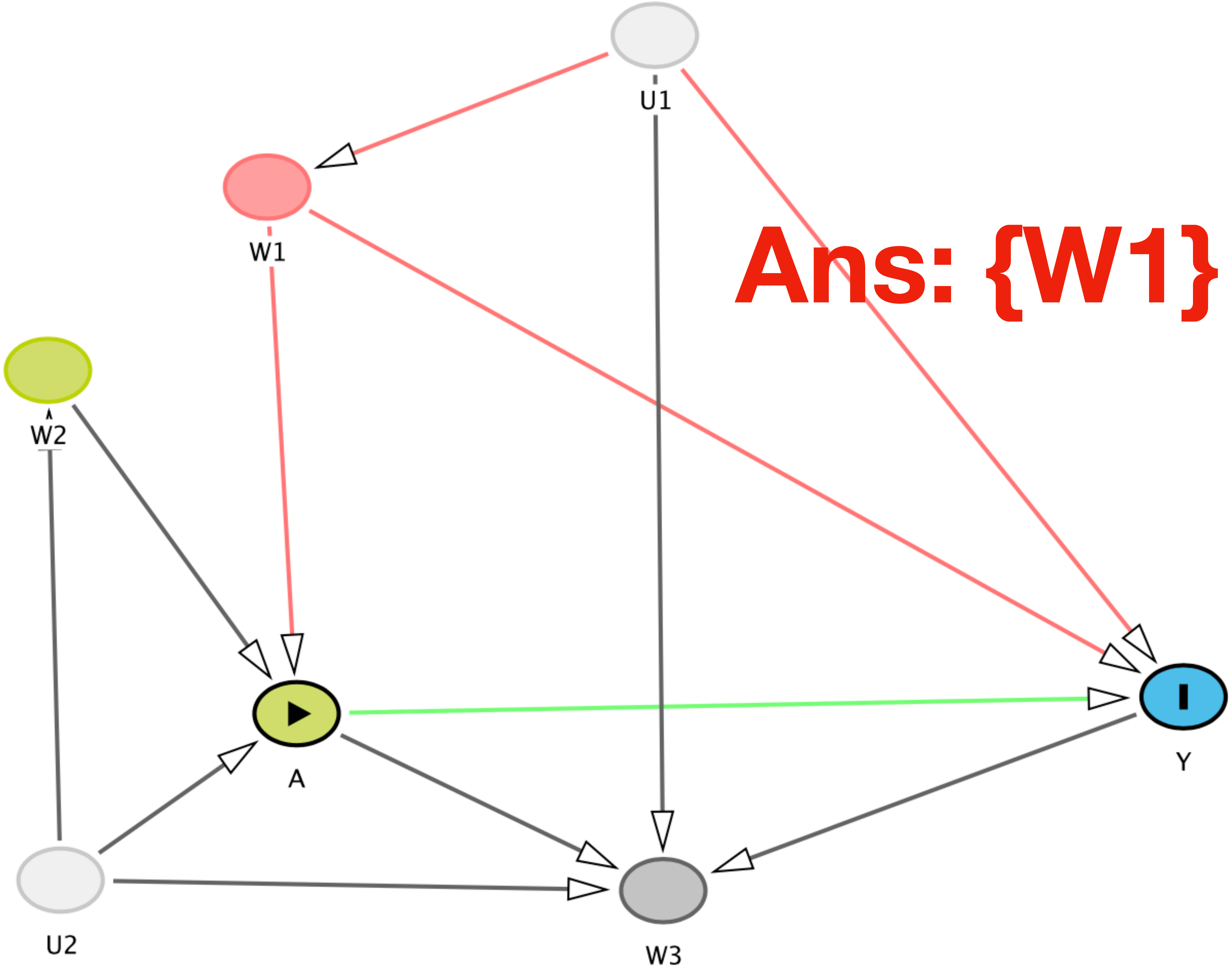
Type your answer in the chat!

What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



Type your answer in the chat!

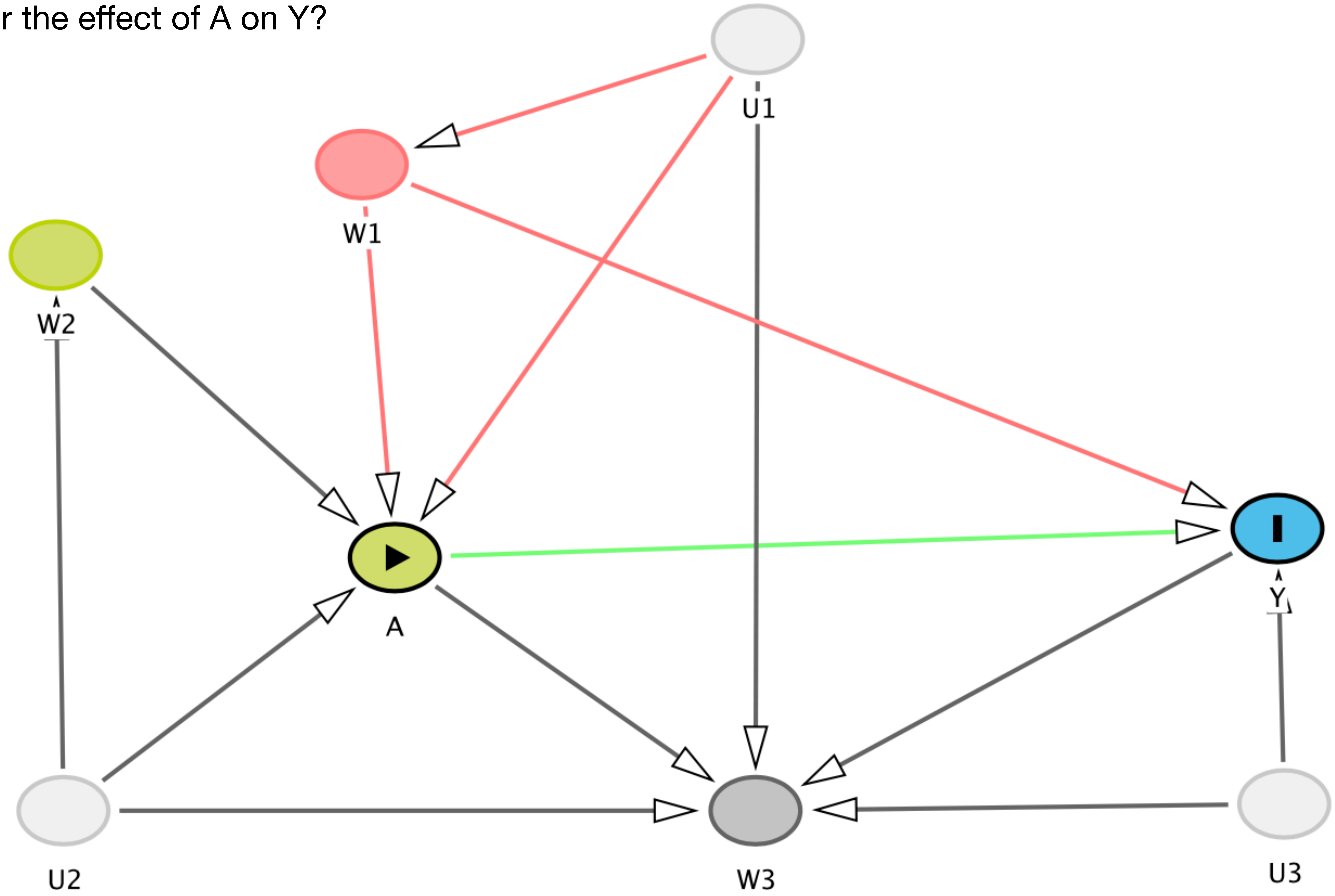
What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



Ans: {W1}

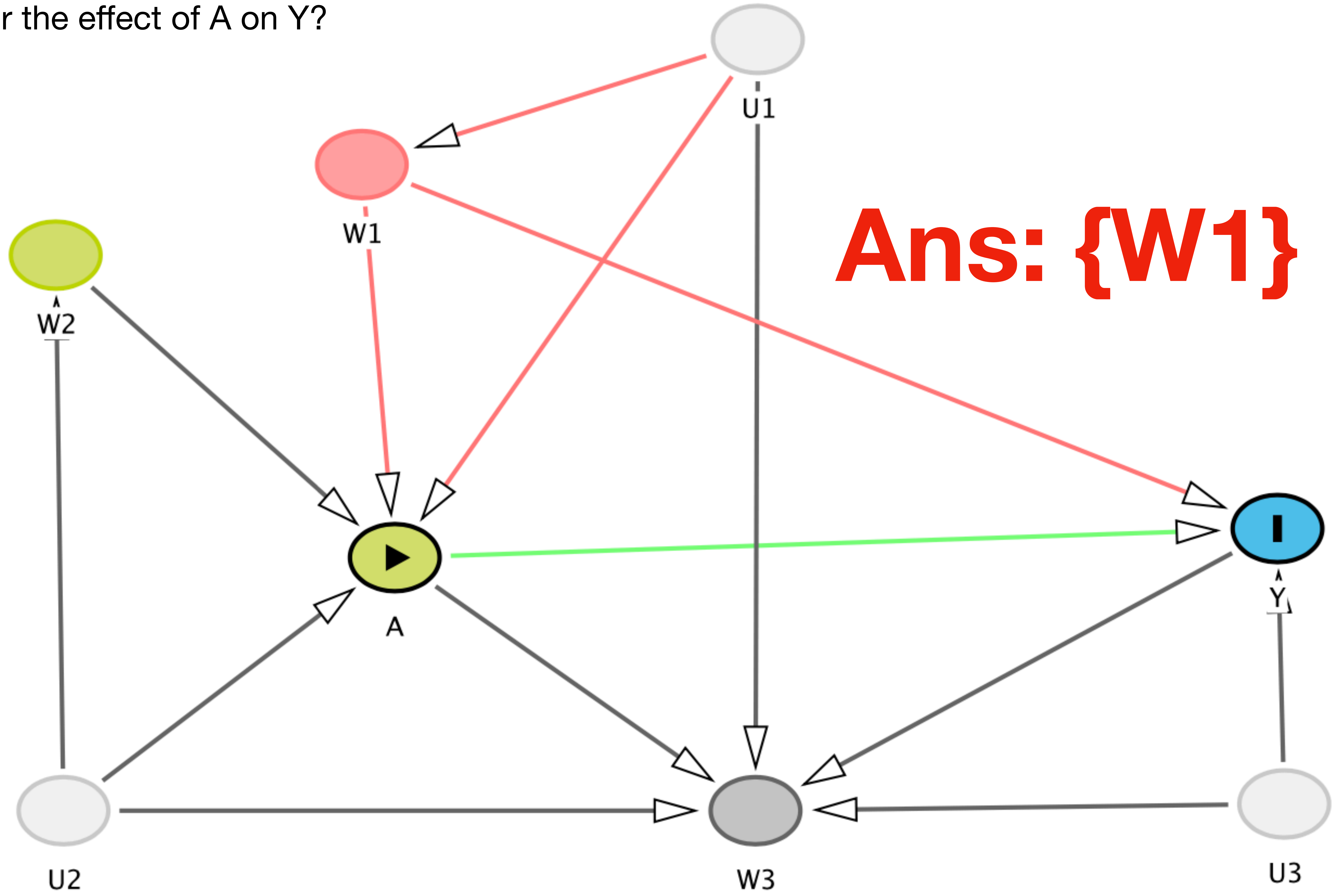
Type your answer in the chat!

What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



Type your answer in the chat!

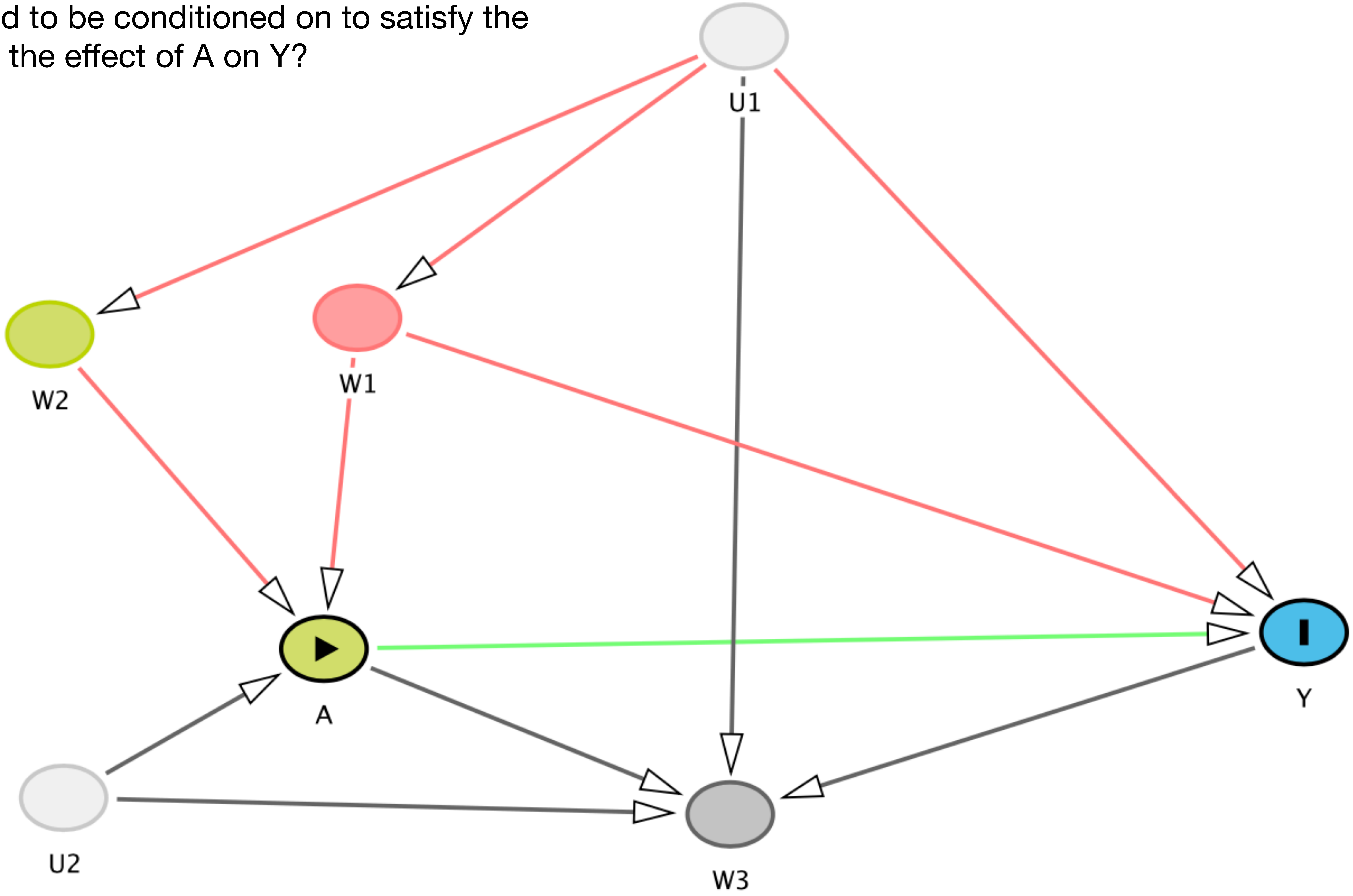
What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



Ans: {W1}

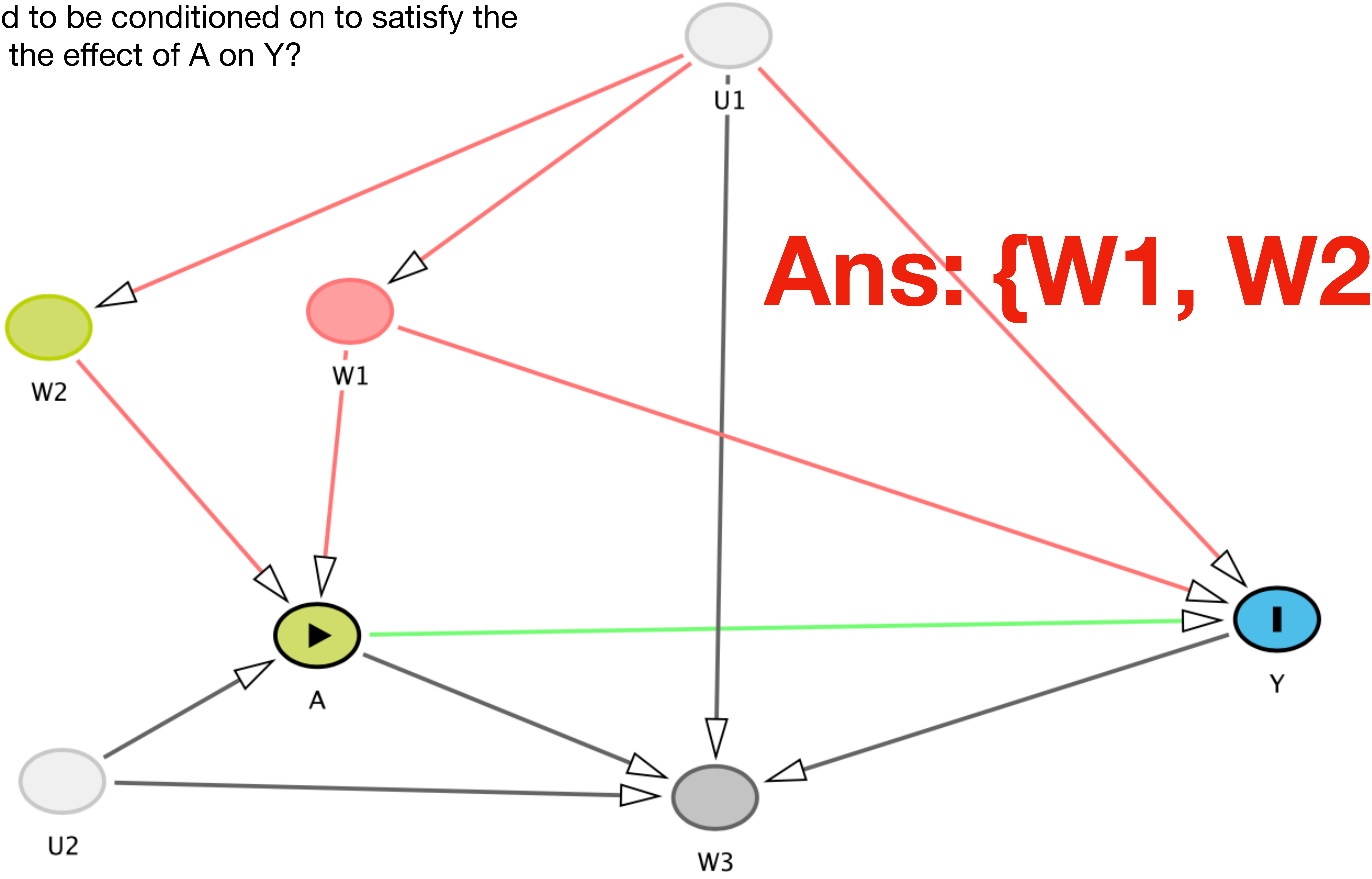
Type your answer in the chat!

What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



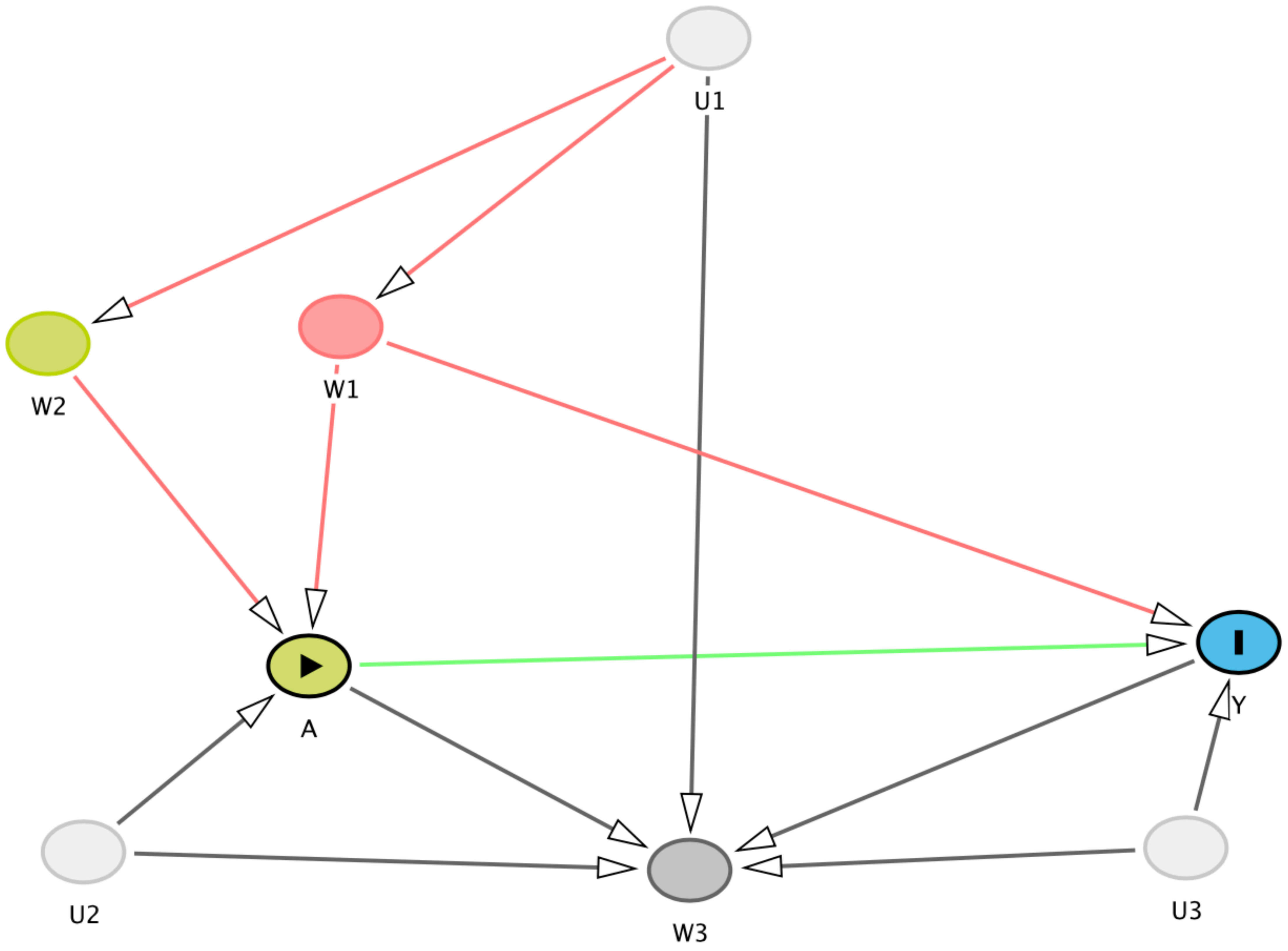
Type your answer in the chat!

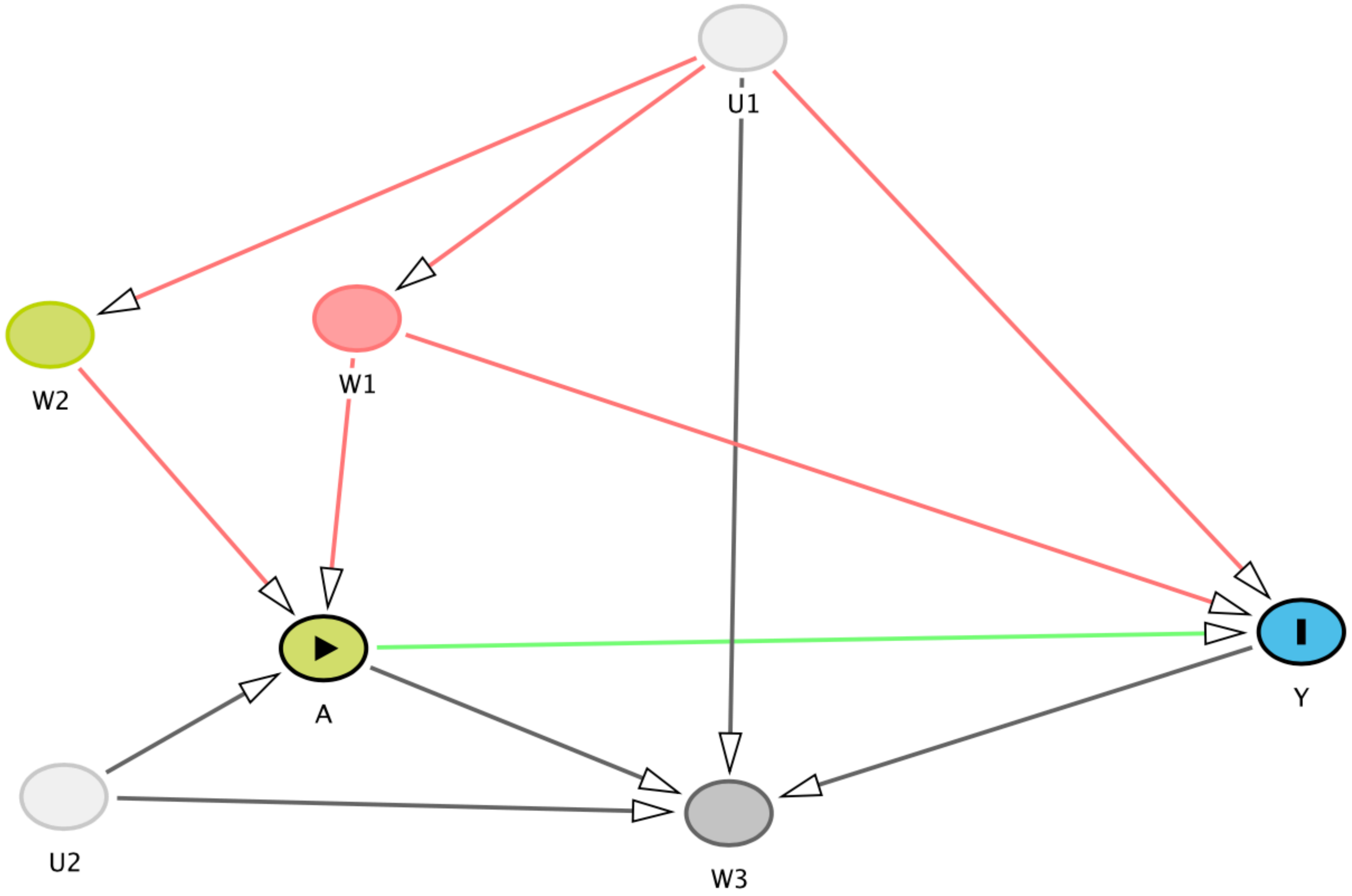
What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



Ans: {W1, W2}

Type your answer in the chat!

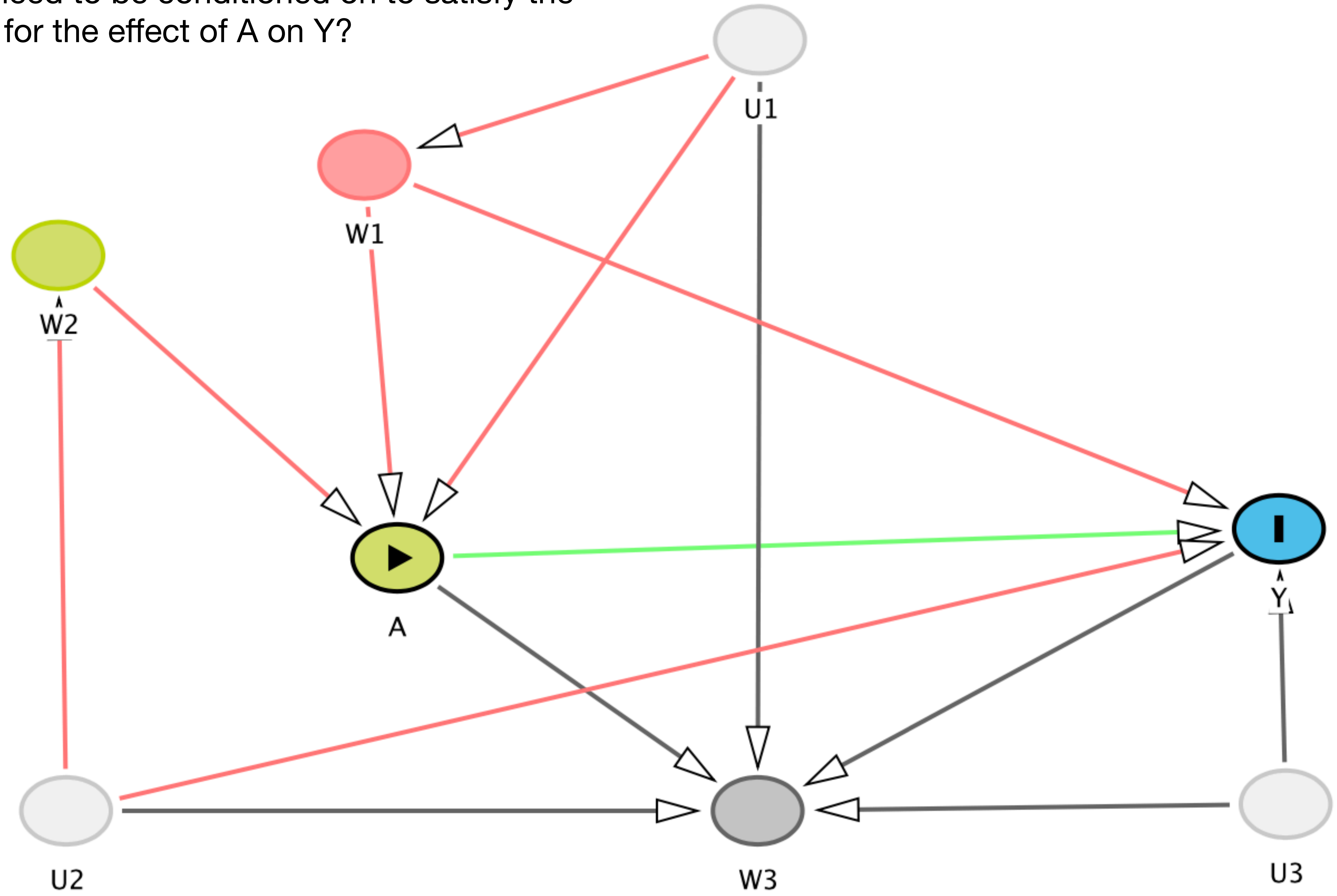




Ans: {W1}

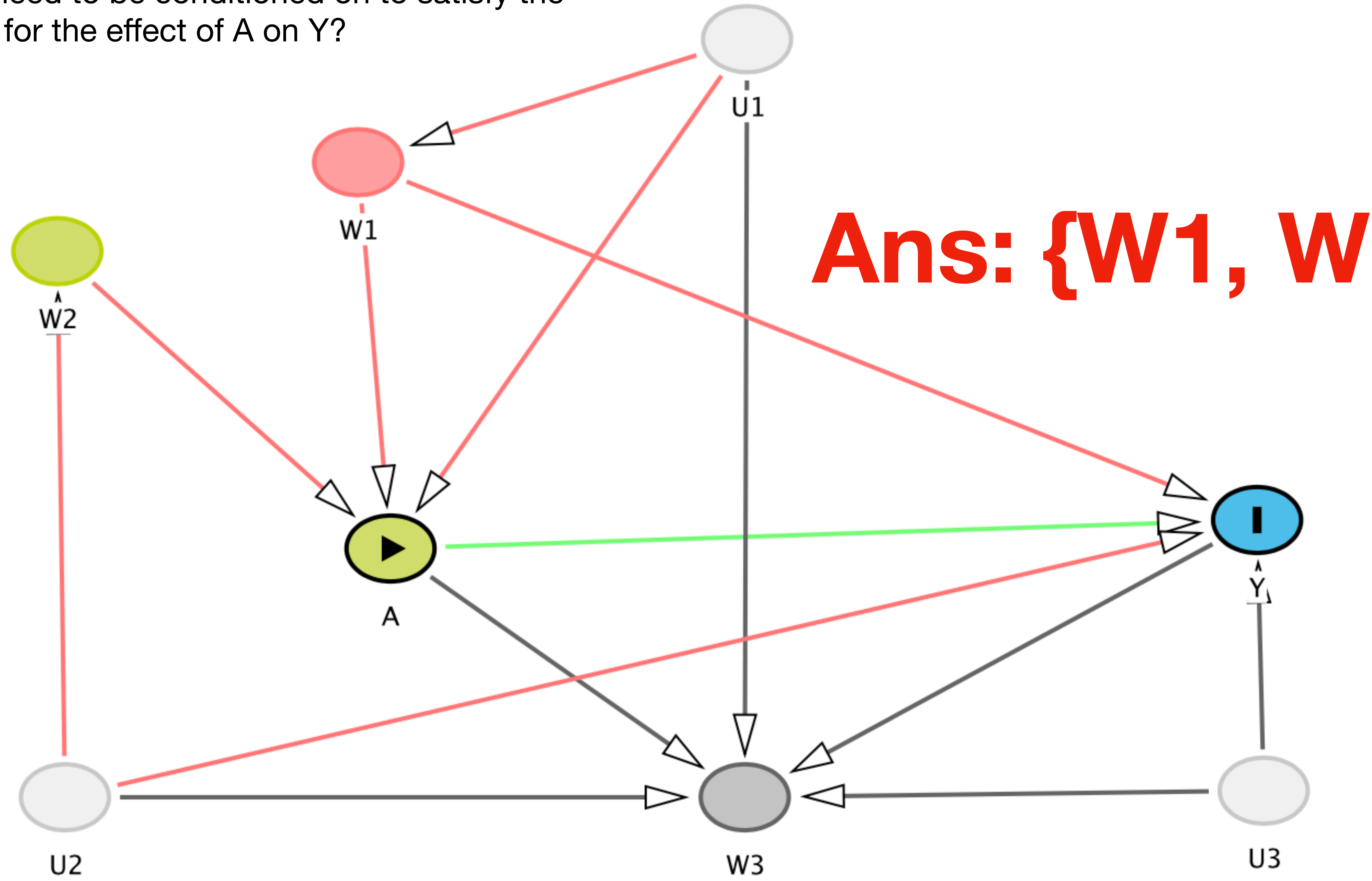
Ans: {W1, W2}

What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



Type your answer in the chat!

What set of nodes need to be conditioned on to satisfy the back-door criterion for the effect of A on Y?



Ans: {W1, W2}

Type your answer in the chat!

“[The Hunger Games] is written in the voice of sixteen-year-old Katniss Everdeen, who lives in a post-apocalyptic world in the country of Panem where the countries of North America once existed. The Capitol, a highly advanced metropolis, holds hegemony over the rest of the nation. The Hunger Games are an annual event in which one boy and one girl aged 12 to 18 from each of the 12 districts surrounding the Capitol are selected by lottery [as ‘tributes’] to compete in a televised battle in which only one person can survive.” - Source: Wikipedia “The Hunger Games”

Some of the tributes have trained extensively for this tournament. The life experiences of other tributes have resulted in certain abilities/advantages (e.g. strength, tree climbing, markmanship). Prior to the tournament, a committee of judges assigns a score to each the tribute indicating his/her probability of winning. Once the tournament starts, forming alliances and sponsorship can aid in survival. A lone victor returns to their district and is showered with wealth and other resources.

Suppose we are interested in the effect of forming an alliance on the probability of surviving through the first 24 hours. We have randomly sampled one tribute from each year of the games. Let $W1$ denote the tribute’s sex with $W1 = 1$ being male and $W1 = 0$ female. Let $W2$ denote the score from the judges. Let A be an indicator that an alliance is formed, and Y be an indicator of survival through the first 24 hours. Finally, let $W3$ be an indicator of whether the tribute receives aid from sponsors during the tournament. Our goal is to evaluate the effect of forming an alliance on the probability of surviving through the first 24 hours.

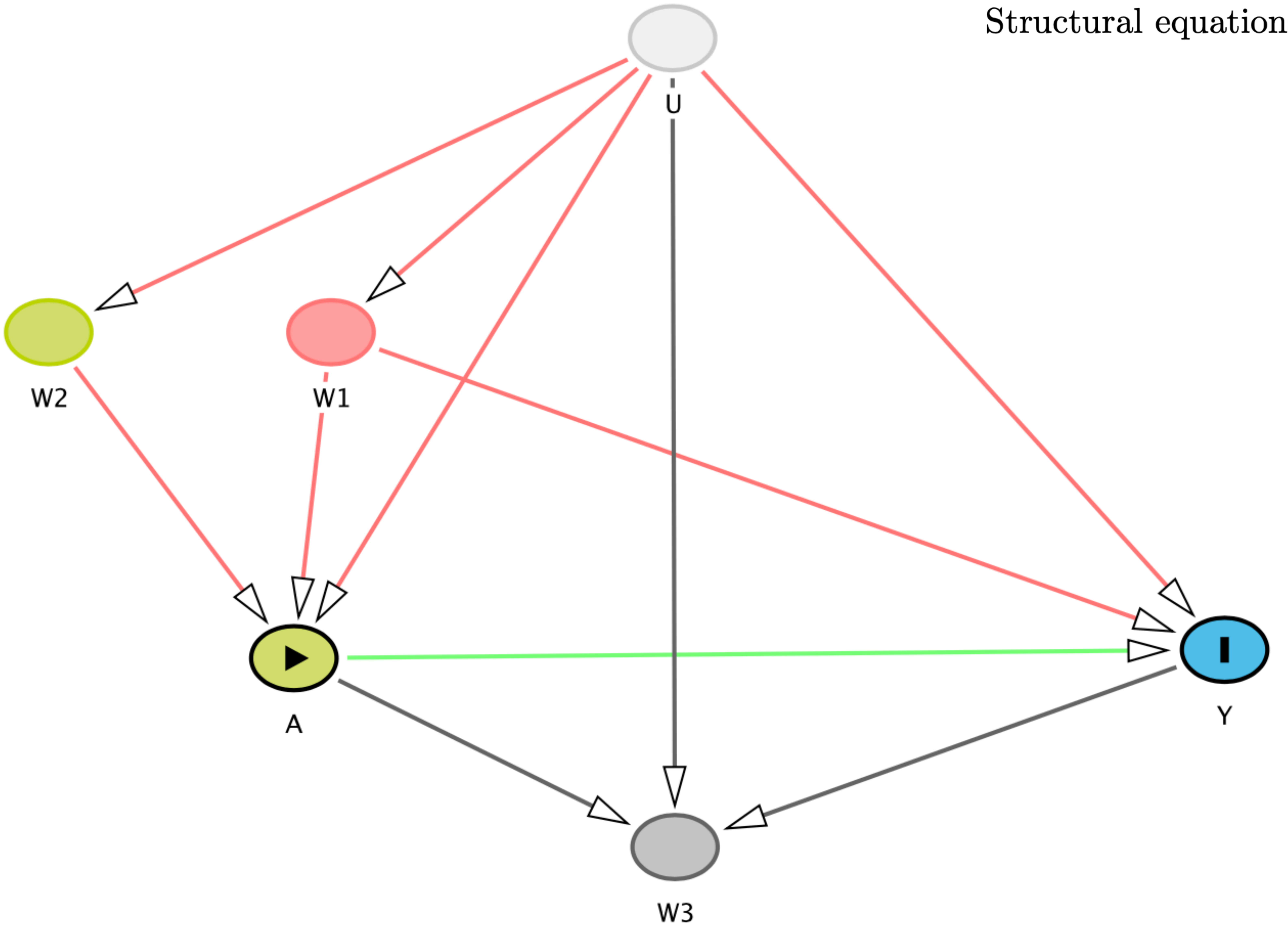
Link O to causal model

We assume the observed data $O = (W1, W2, A, Y, W3)$ were generated by sampling n i.i.d. times from a data generating system compatible with $\mathcal{M}^*$. This provides a link between the causal model $\mathcal{M}^*$ and the observed data O. The distribution of the background variables U and the structural equations F identify the distribution of the endogenous variables X and thus the distribution of the observed data O. We have not placed any restrictions on the statistical model $\mathcal{M}$, which is thereby non-parametric.

Endogenous variables: $X = (W1, W2, A, Y, W3)$
Background variables: $U = (U_{W1}, U_{W2}, U_A, U_Y, U_{W3}) \sim \mathbb{P}_U$
Structural equations F:

$$\begin{aligned} W1 &= f_{W1}(U_{W1}) \\ W2 &= f_{W2}(U_{W2}) \\ A &= f_A(W1, W2, U_A) \\ Y &= f_Y(W1, A, U_Y) \\ W3 &= f_{W3}(A, Y, U_{W3}) \end{aligned}$$

Note: “Restrictions” defined in terms of:
(i) form of structural equations *f*
(ii) allowed distributions for *U* (incl independence)



Endogenous variables: $X = (W1, W2, A, Y, W3)$
Background variables: $U = (U_{W1}, U_{W2}, U_A, U_Y, U_{W3}) \sim \mathbb{P}_U$
Structural equations F:

$$\begin{aligned} W1 &= f_{W1}(U_{W1}) \\ W2 &= f_{W2}(U_{W2}) \\ A &= f_A(W1, W2, U_A) \\ Y &= f_Y(W1, A, U_Y) \\ W3 &= f_{W3}(A, Y, U_{W3}) \end{aligned}$$

Adjustment set related to independence assumptions

For now only condition on W1… which means that our statistical estimand = causal parameter under certain assumptions.
Conditioning on W1 and W2 means statistical estimand = causal parameter under different assumptions. We get to choose if either seems more reasonable.

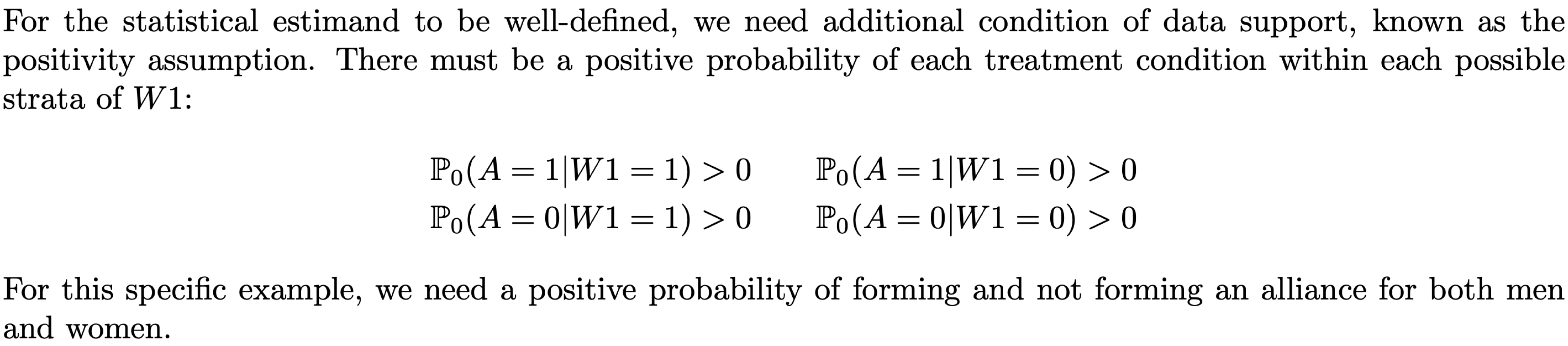
Despite lack of identifiability, we can still "commit" to an interesting statistical estimand inspired by our scientific/causal question:

$$\begin{aligned}\Psi(\mathbb{P}_0) &= \mathbb{E}_0\big[\mathbb{E}_0(Y|A=1,W1) - \mathbb{E}_0(Y|A=0,W1)\big] \\ &= \sum_{w1}\big[\mathbb{E}_0(Y|A=1,W1=w1) - \mathbb{E}_0(Y|A=0,W1=w1)\big]\mathbb{P}_0(W1=w1)\end{aligned}$$

still non-parametric

Suppose we are interested in the effect of forming an alliance on the probability of surviving through the first 24 hours. We have randomly sampled one tribute from each year of the games. Let $W1$ denote the tribute's sex with $W1 = 1$ being male and $W1 = 0$ female. Let $W2$ denote the score from the judges. Let A be an indicator that an alliance is formed, and Y be an indicator of survival through the first 24 hours. Finally, let $W3$ be an indicator of whether the tribute receives aid from sponsors during the tournament. Our goal is to evaluate the effect of forming an alliance on the probability of surviving through the first 24 hours.

What is the positivity assumption (in words)?



For the statistical estimand to be well-defined, we need additional condition of data support, known as the positivity assumption. There must be a positive probability of each treatment condition within each possible strata of $W1$:

$$\begin{array}{ll} \mathbb{P}_0(A=1|W1=1)>0 & \mathbb{P}_0(A=1|W1=0)>0 \\ \mathbb{P}_0(A=0|W1=1)>0 & \mathbb{P}_0(A=0|W1=0)>0 \end{array}$$

For this specific example, we need a positive probability of forming and not forming an alliance for both men and women.

Note:

1) **Outcome Y does not matter here**
2) **If we choose to condition on W1 and W2, this would change.**

G-Computation

one way to estimate our target parameter
(in this case the ATE)

G-Computation

$$\hat{\Psi}(\mathbb{P}_n) = \sum_w \left[\hat{\mathbb{E}}(Y|A=1, W=w) - \hat{\mathbb{E}}(Y|A=0, W=w)\right]\hat{\mathbb{P}}(W=w)$$

where W denotes our adjustment set. We will always use the sample proportion to estimate the covariate distribution $\mathbb{P}_0(W)$ and therefore can express our simple substitution estimator as

$$\begin{aligned}\hat{\Psi}(\mathbb{P}_n) &= \sum_w \left[\hat{\mathbb{E}}(Y|A=1, W=w) - \hat{\mathbb{E}}(Y|A=0, W=w)\right] \times \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(W_i = w) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_w \left[\hat{\mathbb{E}}(Y|A=1, W=w) - \hat{\mathbb{E}}(Y|A=0, W=w)\right] \times \mathbb{I}(W_i = w) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[\hat{\mathbb{E}}(Y|A=1, W_i) - \hat{\mathbb{E}}(Y|A=0, W_i)\right]\end{aligned}$$

where $\mathbb{P}_n$ denotes the empirical distribution, which puts weight $1/n$ on each copy $O_i,\ i = 1, \ldots, n$.

G-Computation

- Big idea:
  - Estimate the mean of Y given A, W… requires thought $\mathbb{E}_0(Y|A,W)$
  - Estimate the distribution of W… easy-ish
  - Generate the difference Y(A = 1) - Y(A = 0) for each strata of W… easy once you have the mean function from step 1
  - Take the expectation over the distribution of W (a.k.a. a weighted mean based on W) to get ATE… easy

Conditional mean function

$$\mathbb{E}_0(Y|A,W)$$

- *Estimate the mean of Y given A, W… requires thought*

What kind of model/function? (Linear, logistic, ML? Anything that can predict…) What specification? (Especially for linear models and GLMs.) Interactions? How to choose?

Later we will see how to jump over many of these problems with Super Learner…

No matter what you choose, we can evaluate our conditional mean function using standard statistical properties: Bias, variance, mean squared error (MSE) using simulations.

(Note: In the R Lab 2 case, it’s actually easy to make an outcome regression equivalent to the mean in each strata of W1.)

B / V / MSE review

Over many attempts to estimate a true value…

- Bias: Is the mean of my estimates close to the truth? `mean(ests - trueval)`

- Variance: How much variability is there in the estimates? `var(ests)`

- Mean Squared Error (MSE): In the name… `mean((ests - trueval)^2)`

Note: MSE = bias^2 + variance

B / V / MSE and G-Comp

If your conditional mean function is good, you should have low bias / variance / MSE when using G-comp to estimate the ATE.

There may be some situations where you will trade off bias for variance etc… not going to go into a lot of detail here.

If you have several competing estimators (example: different models you are considering), you can figure out which is best for the simulated data setting.

Glossing over a LOT of details here (bias-variance tradeoff, how to get ATE inference, what to do when you don’t have a known true value in a real study, problems with overfitting, etc etc etc) … more on that later.

G-Computation in R HW2

1. Calculate true ATE based on a known data-generating process (you did this in R Lab 1 and R HW 1)

2. Generate some data from the DGP

3. Fit a conditional mean model $\mathbb{E}_0(Y|A, W)$ with logistic or linear regression

4. Generate predictions from the model for A = 1 and A = 0 given covariate distribution W

5. Take mean of Y(A = 1) and Y(A = 0) from the predicted values, see how close the estimated ATE is from the true ATE calculated earlier

6. Repeat steps 2-5, measuring bias, variance, MSE across all the estimated ATEs.

G-Computation in R HW2

In R HW 2, you will have four outcome regressions (aka conditional mean models, aka $\mathbb{E}_0(Y|A,W)$ to evaluate. All are logistic regressions.

You will have to examine which is best and comment on whether your results make sense.

Example code for a logistic regression:

```
obs <- gen.data()    #(Assuming you have a data-generating mechanism function already)
reg.model <- glm(Y ~ A + W1 + A*W1, family = 'binomial', data = Obs)    #(Specification of ‘Y ~ …’ may vary)

obs.0 <- obs
obs.0$A <- 0
obs.1 <- obs
obs.1$A <- 1

Y0 <- predict(reg.model, newdata = obs.0, type = ‘response’)
Y1 <- predict(reg.model, newdata = obs.1, type = ‘response’)

estimate.of.ATE <- mean(Y1 - Y0)
```

IMPT REMINDERS

- Josh: stop the recording
- R HW 2 is posted already
- Note that TEMPLATE exists for R HW 2 so that they won't forget any parts of any questions

- I didn't cover EVERYTHING in R Lab 2; please look over the answer key. R code will be very helpful, plus more detail and context